

De-identifying Data

A Way to Safeguard Personal Information

May 24, 2019

SPHPM Registry Special Interest Group

Jacinta Opie & Dianne Brown



People are scared of data breaches...



December 2017

Melbourne University researchers are able to identify individuals based on the "anonymised" data released by Medicare



January 2018

Data released by on-line fitness tracker Strava pinpoints military base in Syria



March 2019

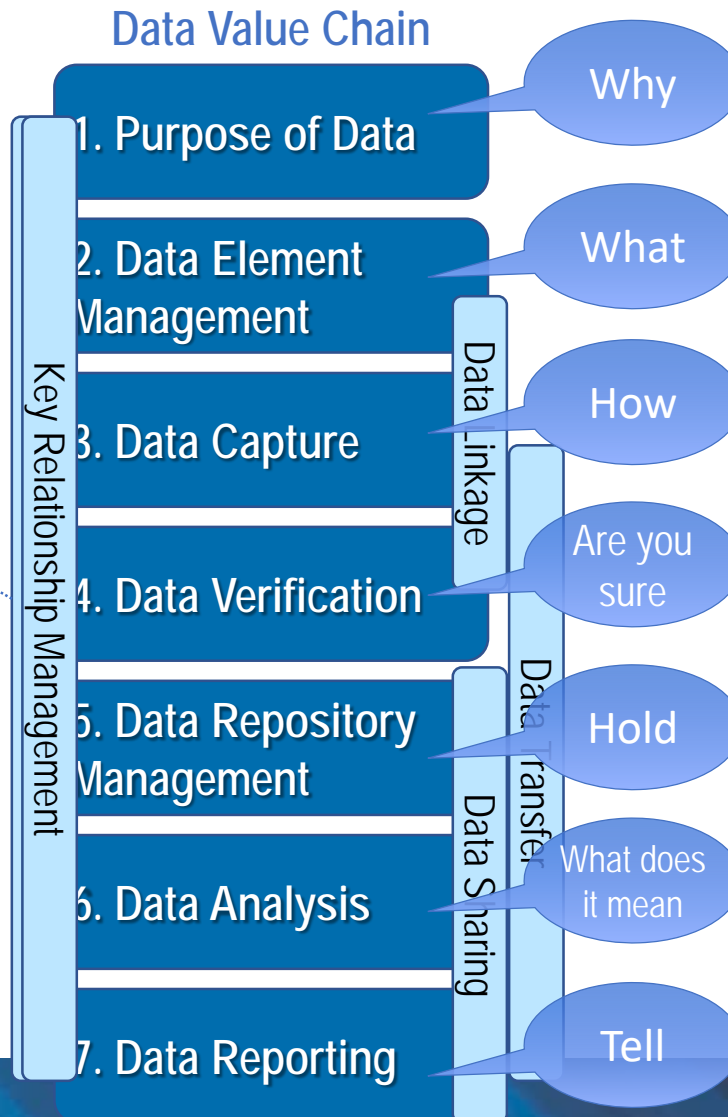
Hackers hold prominent Melbourne Cardiology specialist to ransom after infiltrating their EMR

If people hesitate to share their data, where does that leave research?

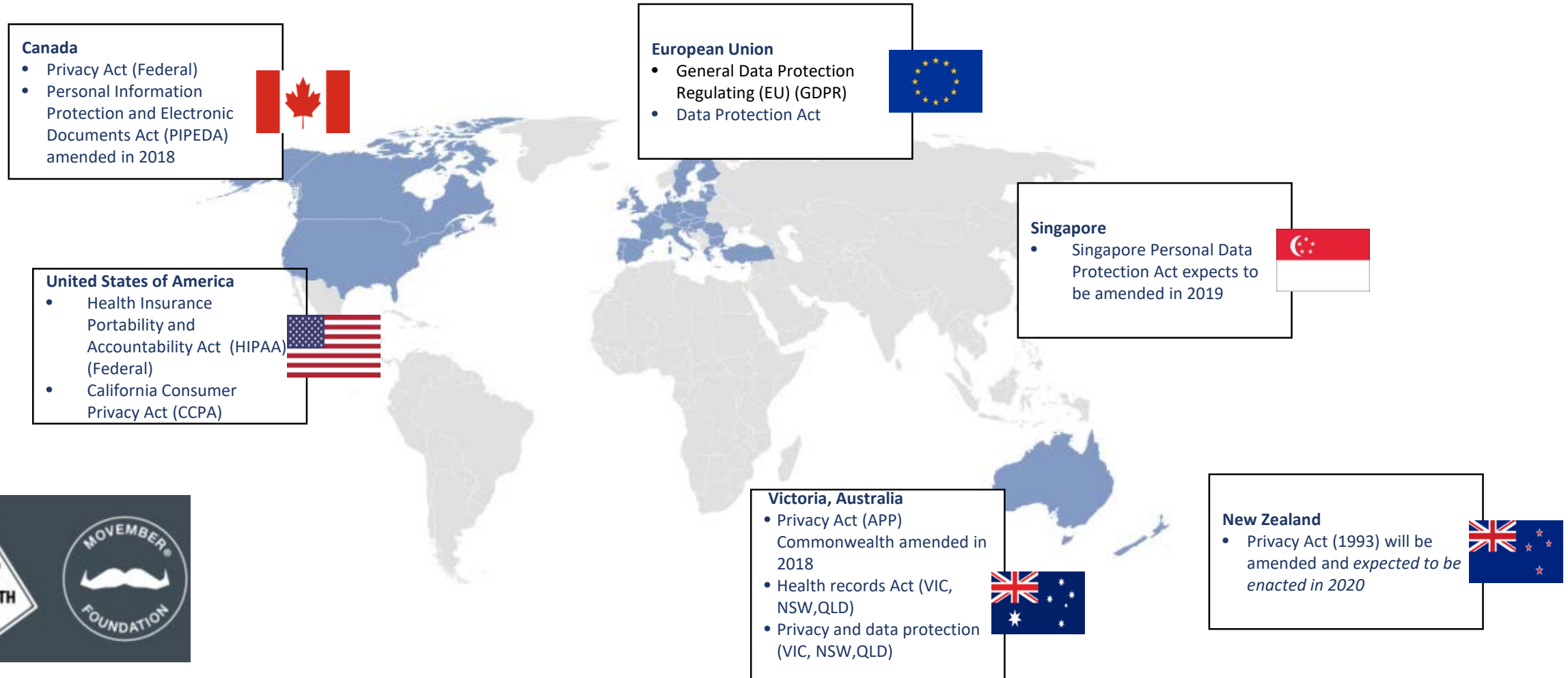
Researchers consider each step of the research process as crucial...

For many participants, the most important step is how their data is protected....

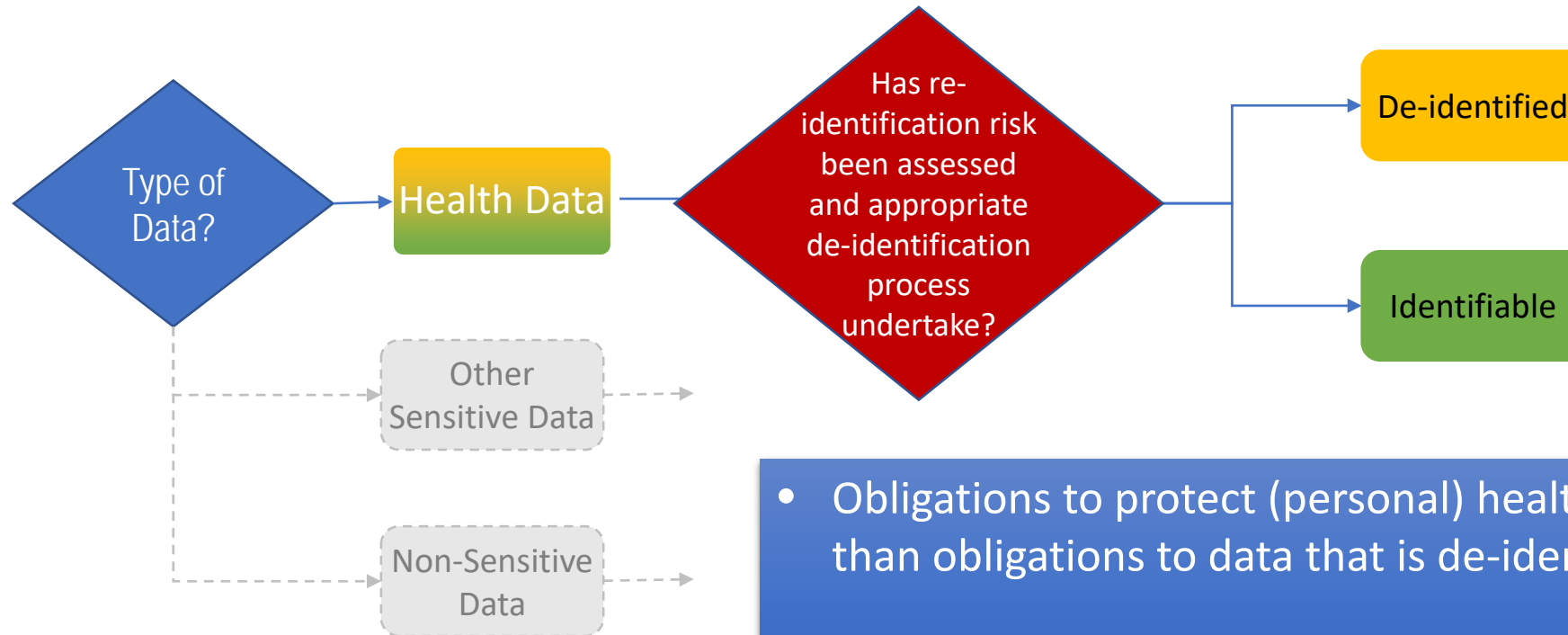
- the measures in place across the value chain to protect the privacy of and provide security for data from unauthorised access



Privacy legislation around the world is changing and complex



So we want to protect the participants' personal data...



- Obligations to protect (personal) health data are much higher than obligations to data that is de-identified
- De-identification safe guards personal information but the risk of re-identification has to be properly considered and assessed

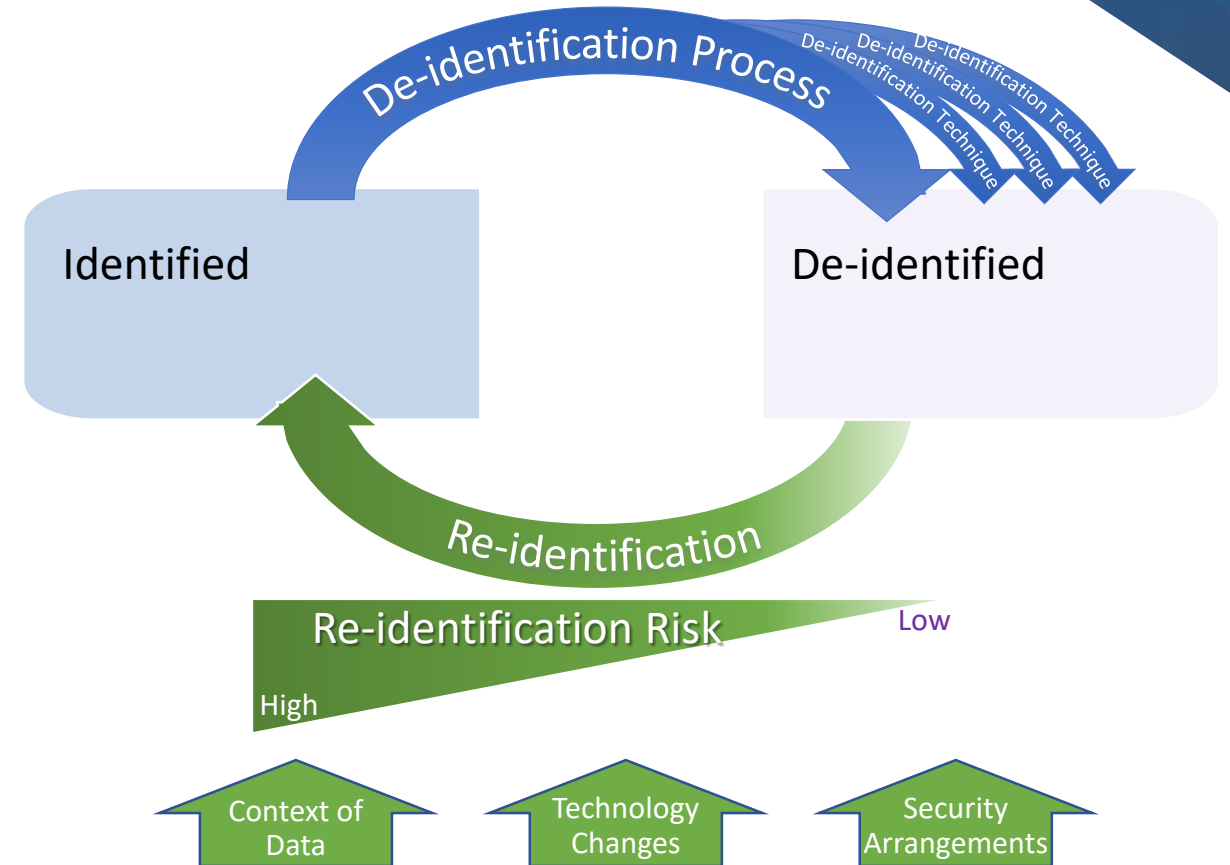
De-identifying data is no longer just a matter of deleting name and date of birth....

Identifiable data becomes **de-identified data** after it undergoes a **de-identification process** (where one or more **de-identification techniques** are undertaken)*.

The appropriate technique(s) to be used will depend on the risk of **re-identification** – ie moving back from de-identified to identified data.

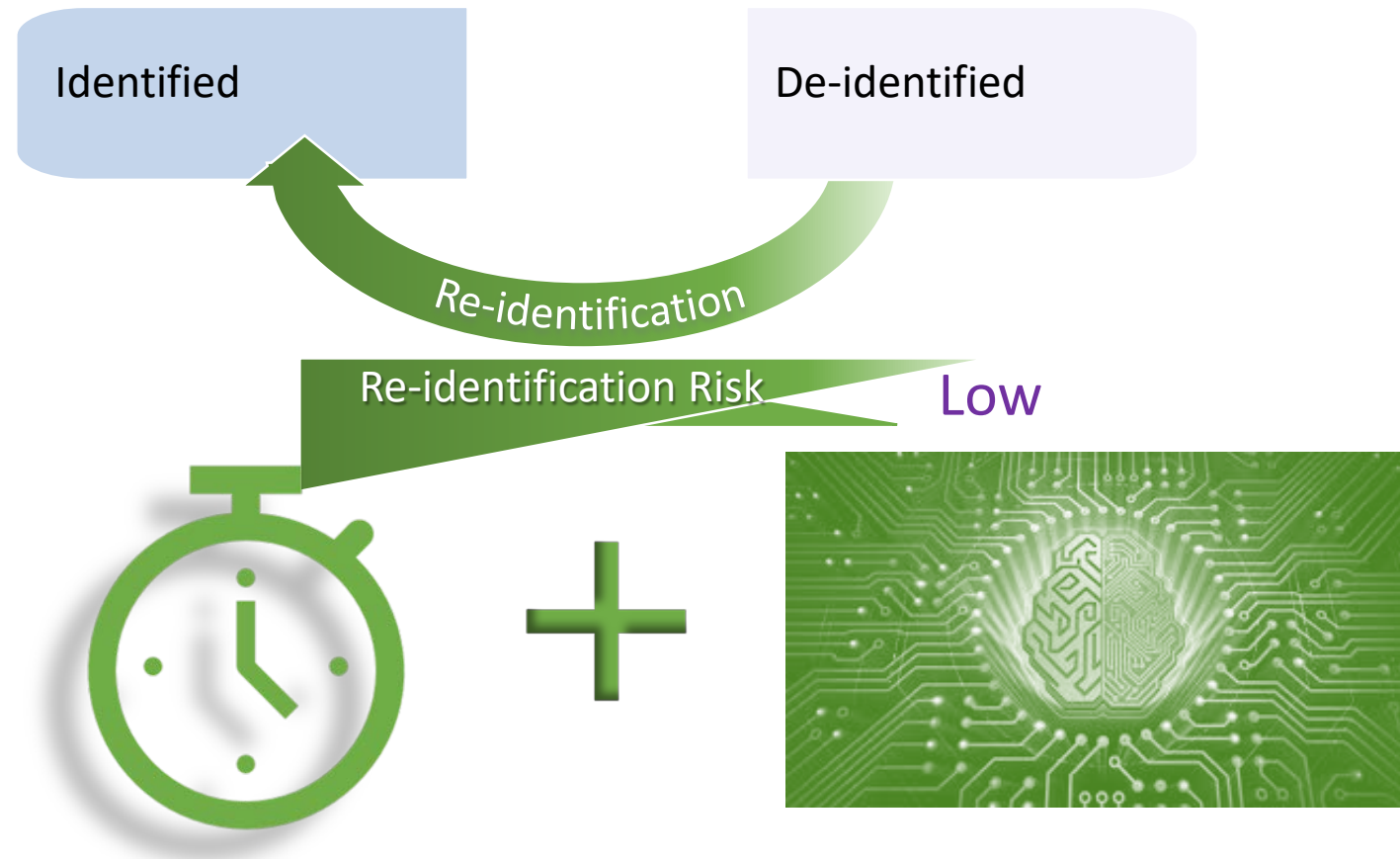
The **re-identification risk** is determined by:

1. Context of the Data
2. Technology Changes
3. Security Arrangements



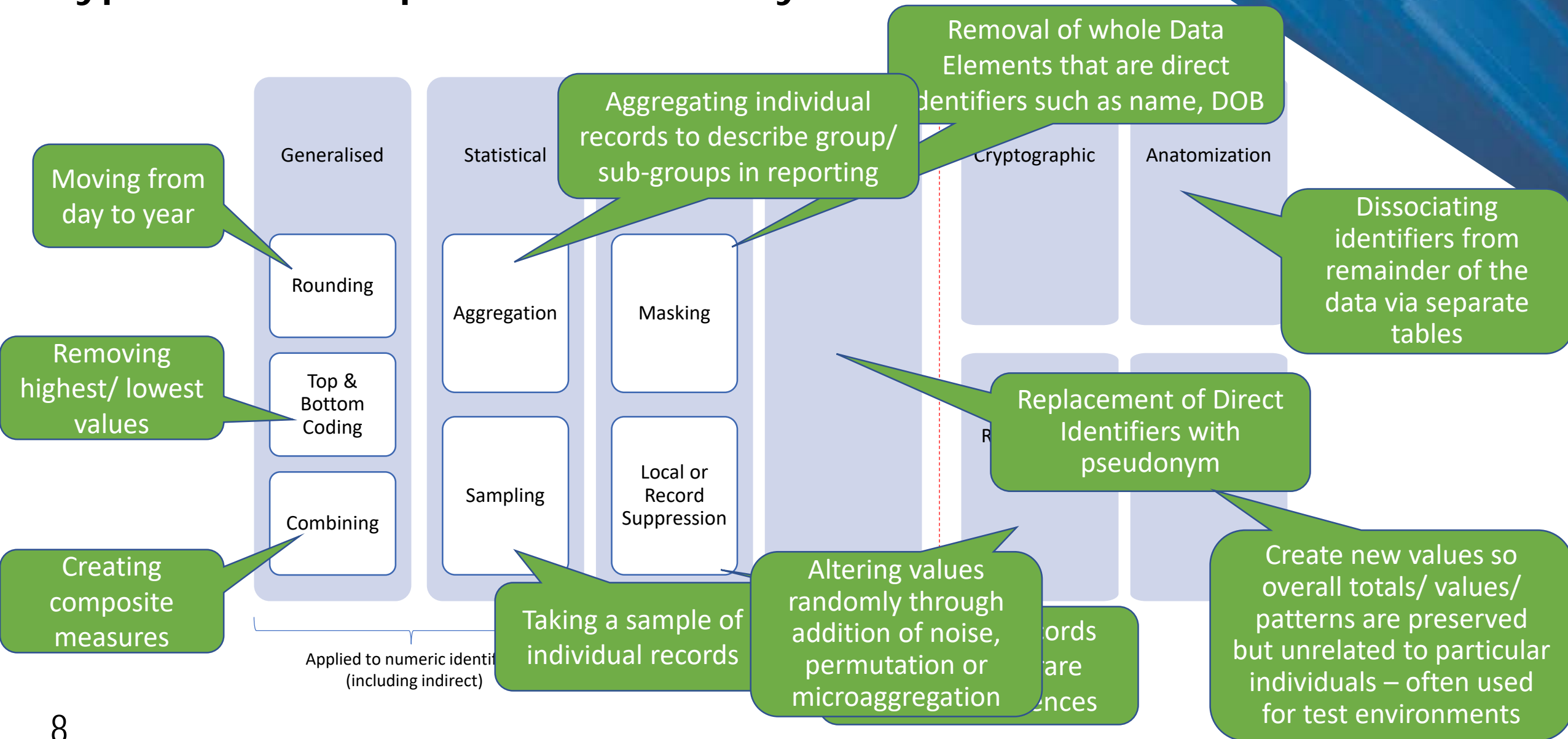
Re-identification risk will never be zero....

With enough computing power and time nearly anything can be re-identified...



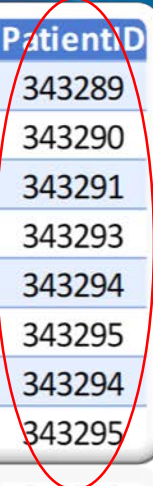
We need to reach a point where an individual can no longer be reasonably identified from the data

Types of techniques available vary....



A word on Pseudonymisation

- **Pseudonymisation** means the processing of personal data so it can no longer be attributed to a specific person without the use of additional information. This additional information must be kept separately and is subject to technical and organisational measures to ensure that the personal data are not used to re-identify an individual.
- In the past pseudonymisation has allowed researchers to hold a key somewhere that would “unlock” the data set and allow re-identification of individuals for data verification purposes, linkage purposes or other data sharing purposes.
- Pseudonymised data will not necessarily be considered de-identified or anonymised

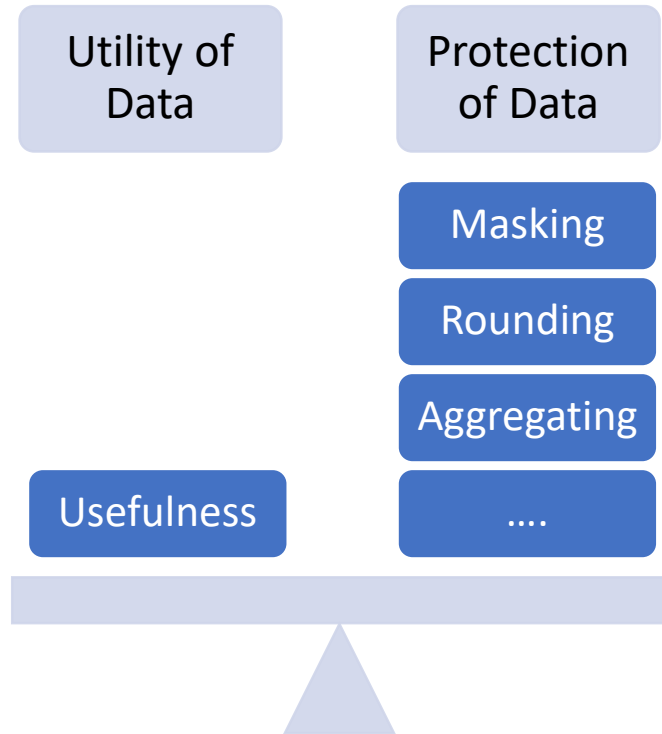


FirstName	Surname	DOB	Address	Suburb	Postcode	PatientID
Maggie	More	7/12/1988	2 Wayride Ave	Brisbane	4000	343289
Raymond	Pulse	1/04/1956	22 Truman Street	Melbourne	3000	343290
Bettina	Kernow	4/07/1978	36 Newman Grove	Sydney	2000	343291
Jacob	Leader	7/09/1954	26 Safety Beach Rd	Perth	8000	343293
Jenny	Brown	7/03/1966	22 Bond St	Adelaide	5000	343294
Marcus	Sparrow	23/05/1967	45 Arterial Rd	Curtin	2132	343295
Olivia	Plum	17/03/1977	4 St Johns Ave	Darwin	8176	343294
Alistair	Beeker	4/07/1978	2/1 High Street	Hobart	2897	343295



If a key exists **anywhere** to re-identify data then it may not be de-identified

Deciding on which technique(s) to use must balance two competing factors....



- The more the data is altered to protect the data through aggregation, masking, suppression, rounding, etc, utility of the data can be lost
- Finding the balance between the two will vary depending on the context of the data which means you must understand the purpose of the data collection in the first place....



Case Study: Bourke Street Attack

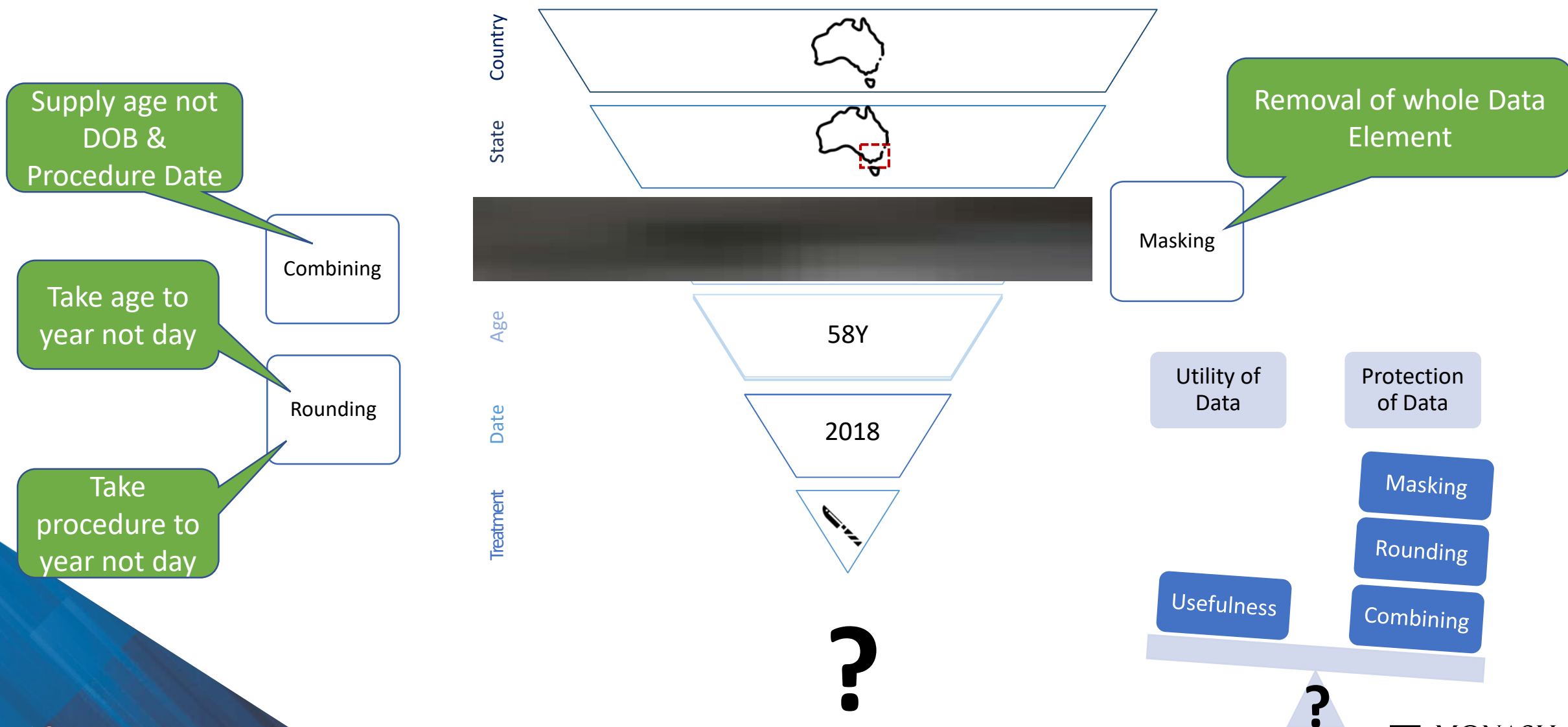


Mr Patterson, 58, suffered head injuries and underwent surgery. He spent Saturday recovering at The Alfred hospital.

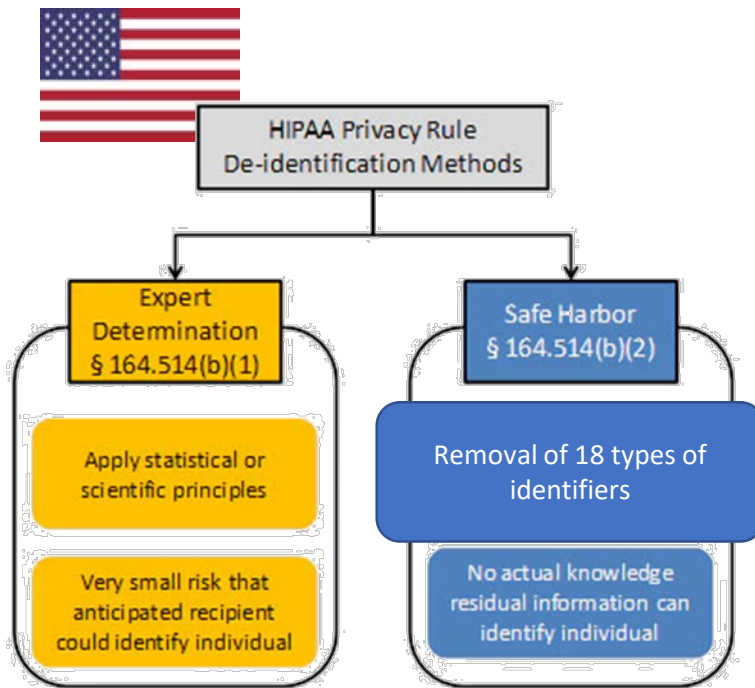


- Alfred Hospital
- Male
- 58
- Head injuries
- Surgery
- 10 Nov 2018

What elements do we need to alter?



Other Jurisdictions provide guidance...



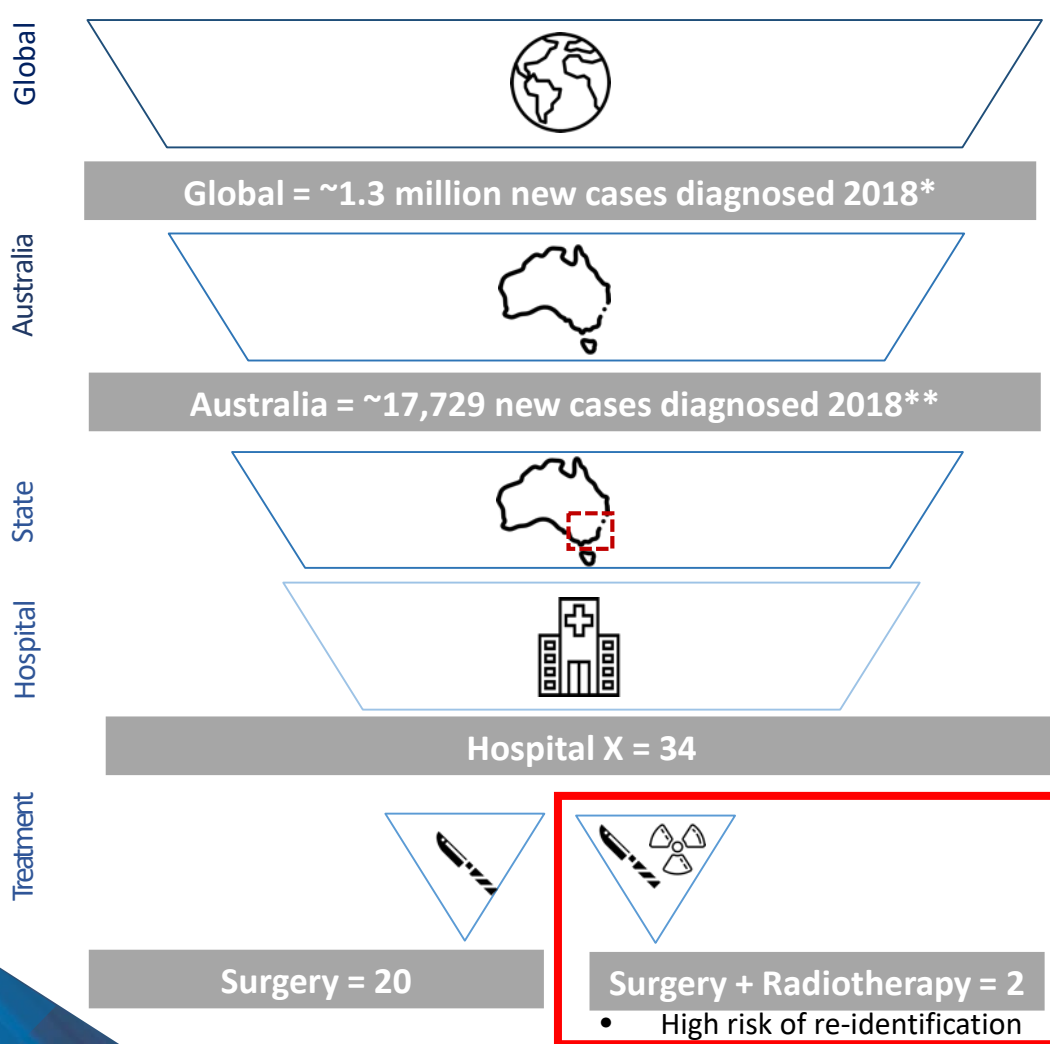
18 Data Elements Considered “Identifiers”

1. Names
2. Geographic location below state level
3. Dates (excluding year)
4. **Ages over 90**
5. Telephone & Fax numbers
6. Vehicle identifiers and serial numbers, including license plate numbers
7. Device identifiers and serial numbers
8. Email addresses
9. Web Universal Resource Locators (URLs)
10. Social security numbers
11. Internet Protocol (IP) addresses
12. Medical record numbers
13. Biometric identifiers, including finger and voice prints
14. Health plan beneficiary numbers
15. Full-face photographs and any comparable images
16. Account numbers
17. Other unique identifying number, characteristic, or code
18. Certificate/license numbers

Top & Bottom Coding

Remove ages over 90 & create category – Over 90

Even aggregation can be a problem when we stratify...



Aggregation

Aggregating individual records to describe group/ sub-groups in report on outcomes in prostate cancer treatment

Selection Data

Combining

Usefulness

Aggregating

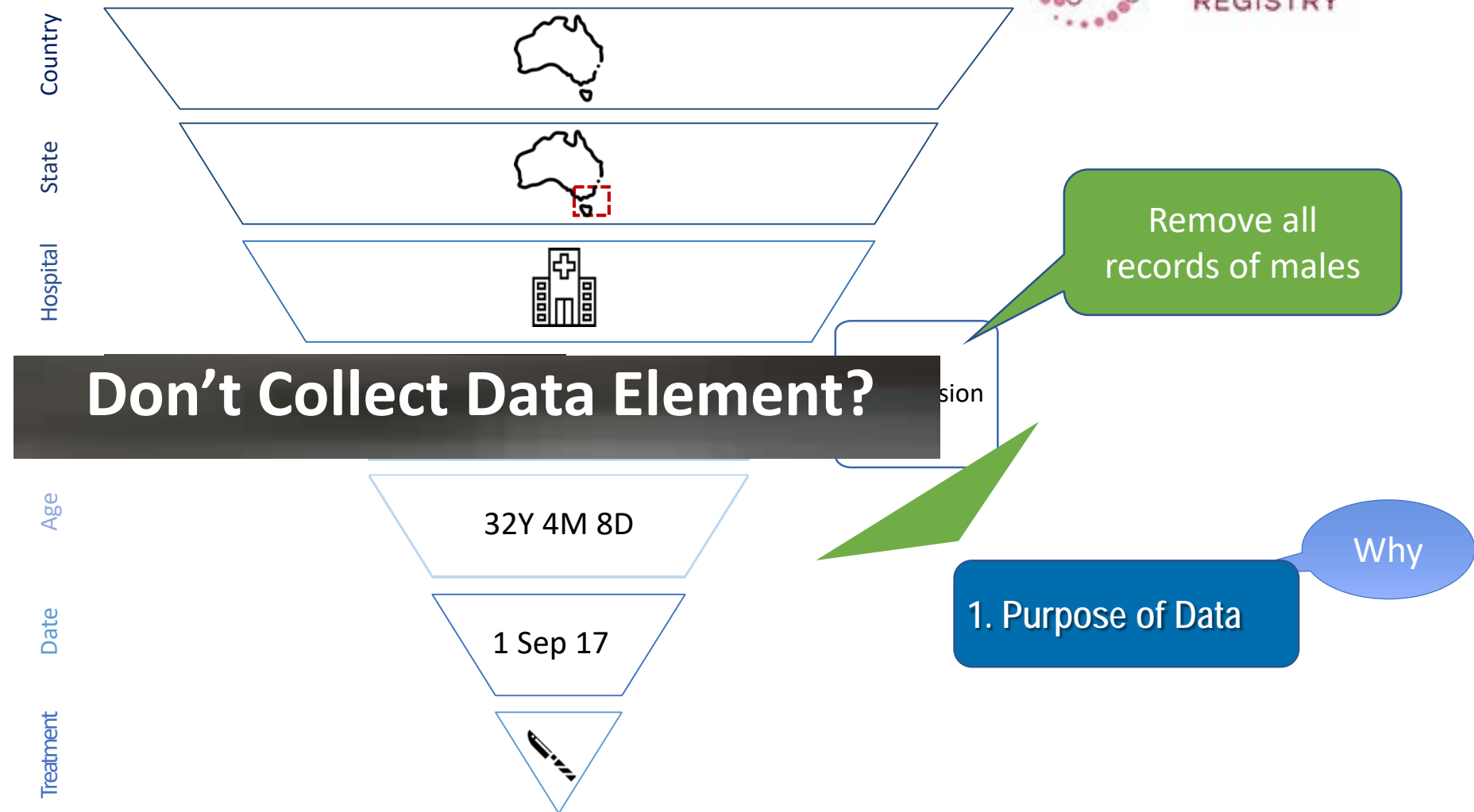
?

Combine

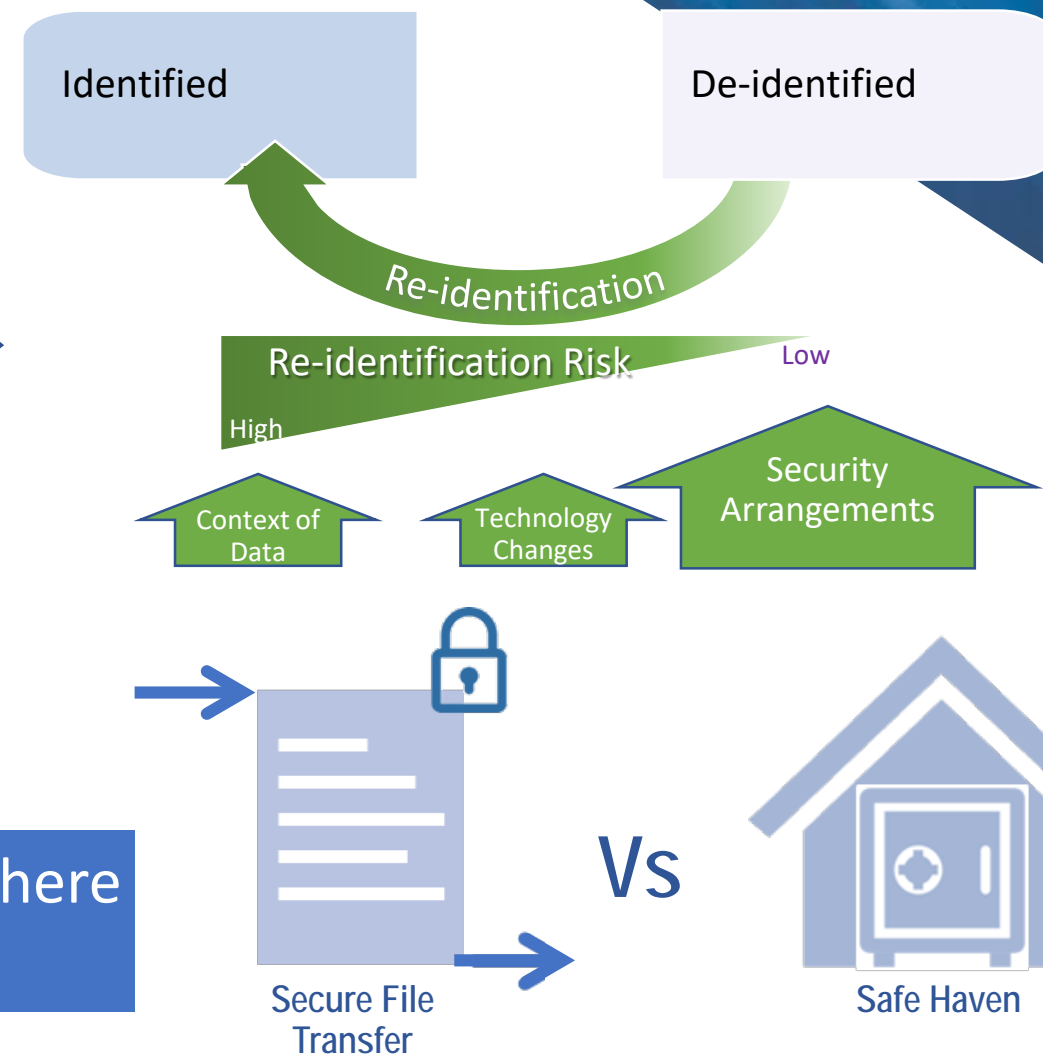
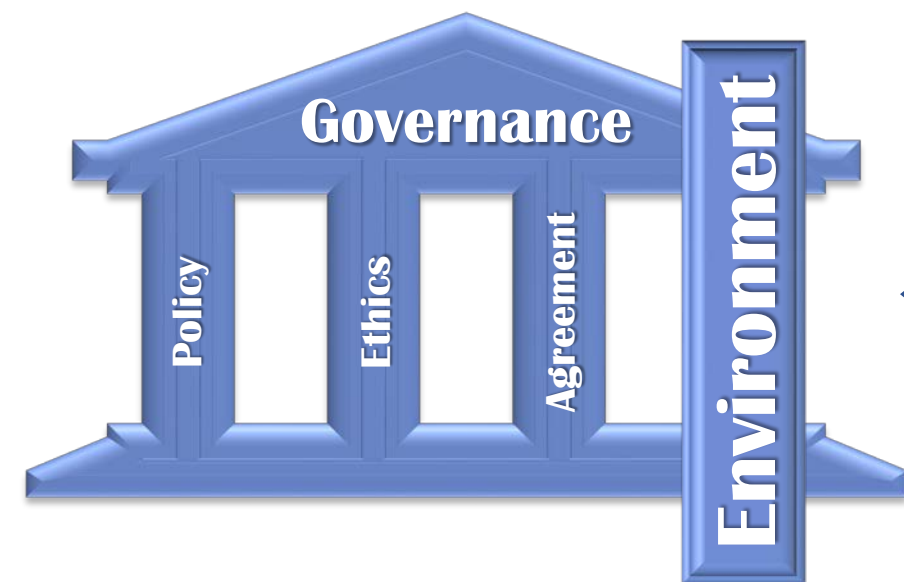
Surgery (with/without) Radiotherapy = 22



Is suppression the best strategy?



Re-identification Risk in Data Sharing



Re-identification risk is lower in a Safe Haven where you can control the linking to other datasets.

De-identification Checklist



Direct Identifiers

Have these all been removed including images?

HIIPA's 18 data elements to remove is a good start



Destination

Where is the data heading? Into the public via Reporting? Data sharing of individual records via file transfer or safe haven?

Think about minimum cell size when stratifying reports...If Data Sharing ensure governance in place & contact Helix about access to sharing environments



Contained

What is actually contained in the data? Are there records that are rare or unusual that make it reasonably likely they can be re-identified?

Look through data and think about Top/ Bottom Coding....individual record suppression....



Pseudonymisation

Do I have all the right measures in place given this may not be recognised as “de-identified” data?

If Data Sharing, do I need to pseudonymise this data or can it be fully de-identified?



Usefulness vs Protection

Given the purpose of my data collection, has the right balance been struck?

Some Useful Resources



<https://www.oaic.gov.au/agencies-and-organisations/guides/de-identification-and-the-privacy-act>



<https://www.oaic.gov.au/agencies-and-organisations/guides/de-identification-decision-making-framework>



<https://ovic.vic.gov.au/wp-content/uploads/2018/08/De-identification-Background-Paper-Update.pdf>



- 18 Data Elements Considered "Identifiers"**
- Names
 - Geographic location below state level
 - Dates (excluding year)
 - Ages over 90
 - Telephone & Fax numbers
 - Vehicle identifiers and serial numbers, including license plate numbers
 - Device identifiers and serial numbers
 - Email addresses
 - Web Universal Resource Locators (URLs)
 - Social security numbers
 - Internet Protocol (IP) addresses
 - Medical record numbers
 - Biometric identifiers, including finger and voice prints
 - Health plan beneficiary numbers
 - Full-face photographs and any comparable images
 - Account numbers
 - Other unique identifying number, characteristic, or code
 - Certificate/license numbers

Acknowledgments

Monash University Data Protection and Privacy Office (DPPO)
<https://www.monash.edu/privacy-monash>

Monash university Data Protection Officer – Susan Anderson
DataProtectionOfficer@monash.edu

KPMG – Stephanie Doidge

Prostate Cancer Registry (PCOR) – Prof Sue Evans, Jade Ting (summer student), Fanny Sampurno

Monash University - eSolutions, HELIX, eResearch teams