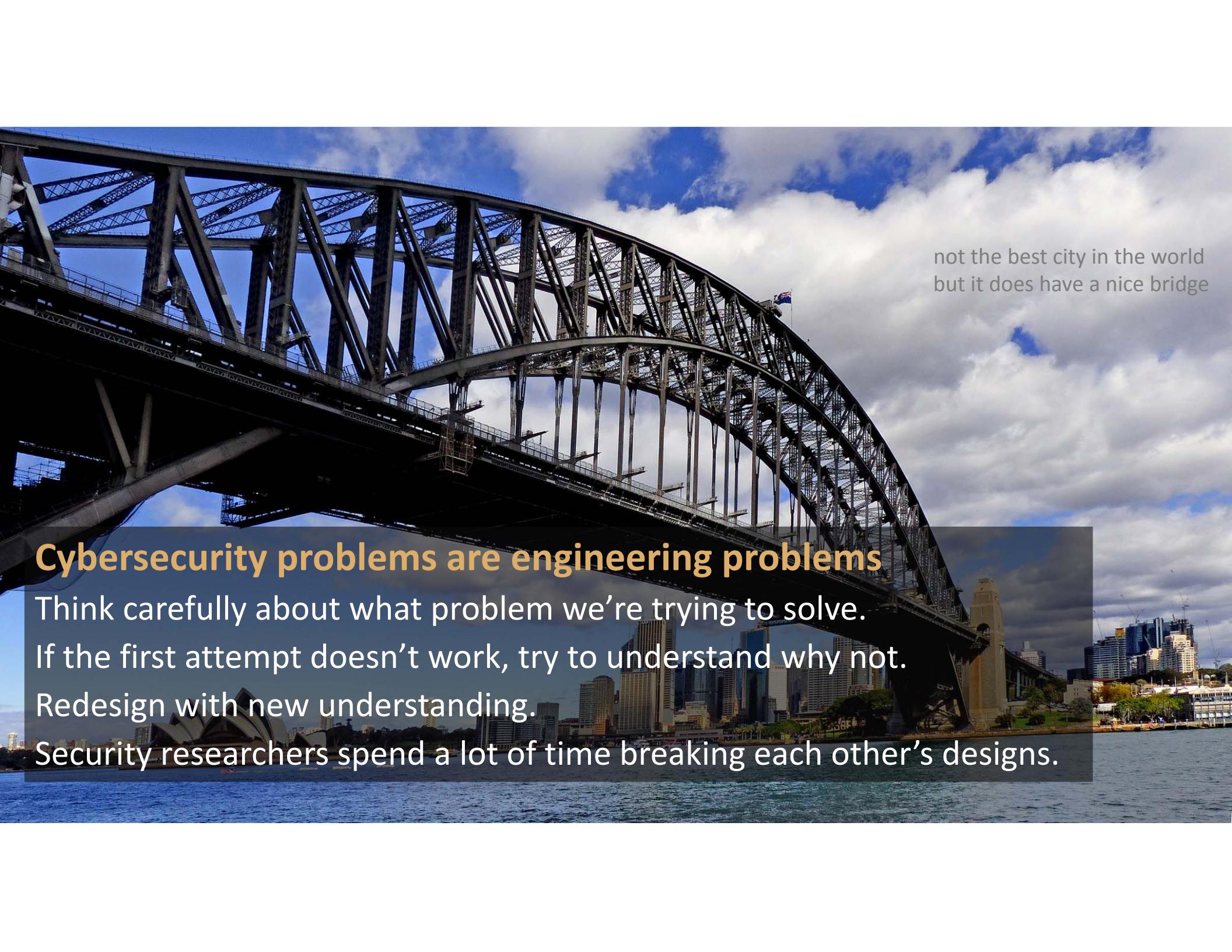




De-Anonymising the Australian Medicare Data Release

Ben Rubinstein with Chris Culnane & Vanessa Teague
The University of Melbourne





not the best city in the world
but it does have a nice bridge

Cybersecurity problems are engineering problems

Think carefully about what problem we're trying to solve.

If the first attempt doesn't work, try to understand why not.

Redesign with new understanding.

Security researchers spend a lot of time breaking each other's designs.

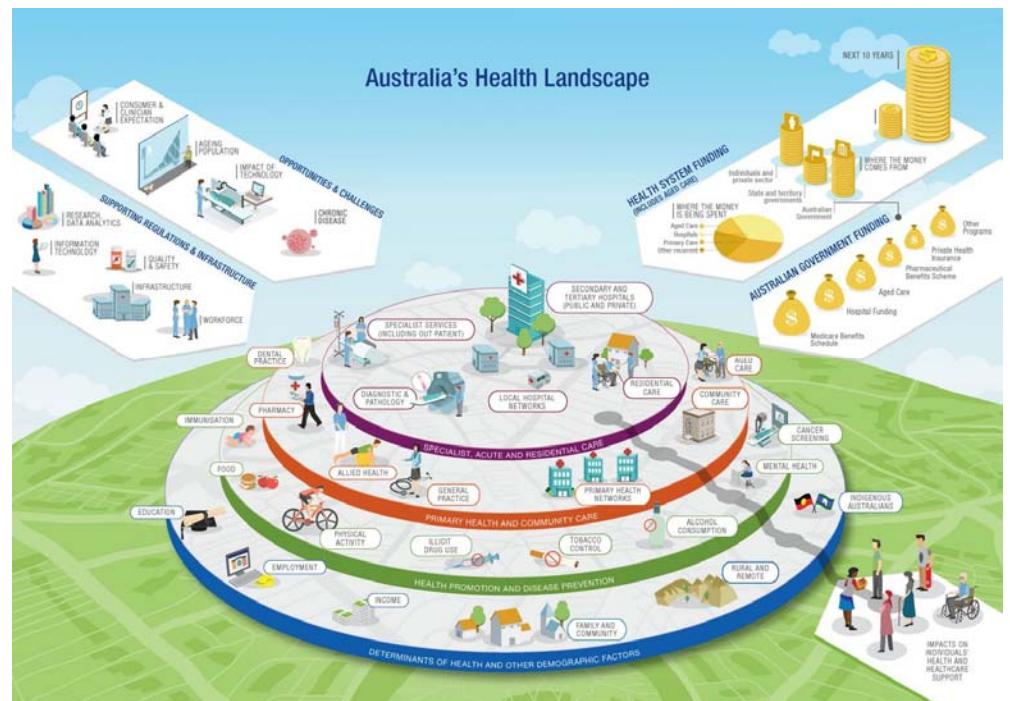
1 August 2016 – MBS/PBS dataset is released

Medicare/Pharmaceutical Benefits Schedules

- *“Medicare is Australia’s universal health insurance scheme... It guarantees all Australians (and some overseas visitors) access to a wide range of health and hospital services at low or no cost.”*
- *“PBS provides timely, reliable and affordable access to necessary medicines for Australians... Under the PBS, the government subsidises the cost of medicine for most medical conditions.”*

3 billion lines of data

- **10%** (2.9 million) of the population
- MBS **30 years**: 1984 – 2014
- PBS 11 years: 2003 – 2014
- Almost*: **Demographics + billing** records



Infographic from the Dept of Health



*What protections were in place?

Encrypted Provider IDs and Patient PINs

Collapsed locations into 4 geographic regions

- patient MBS enrolment
- provider location

Year of birth

Removal of centenarians

Service & supply date perturbed by up to 14 days

- Constant perturbation within a patient's records

Extremely low service volume items removed

“GP” in U.S.: “Family Doc”.
Or: HIV medicine,
2nd trimester labor, ...

(Encrypted) patient ID	0345952108
Gender	F
Year of birth	1963
(Encrypted) patient ID	0345952108
State	Vic-Tas
Date	7 Aug 1992
(Encrypted) supplier ID	2340981234
Item code	00023 (GP visit)
Price paid by patient	\$85
Price reimbursed by Medicare	\$60
Various other details	...



8 September 2016: Decryption reported to DHS

Decrypted Provider IDs in full

- described pseudo RNG insecure in setting
- reversable by guesswork

Aust Privacy Commissioner: Encryption dates back...

- 2005: DHS for 10% PBS releases to select groups
- mid-1990s: former Health Insurance Commission

Instead: should have used RSA, AES, random IDs

We responsibly disclosed to Dept

- providers can be recovered
- high risk of patient reID also

ABC NEWS LOCATION: Melbourne, Vic Change ▾ Cloudy

Just In Politics World Business Sport Science Health Arts Analysis F

Print Email Facebook Twitter More

Medicare dataset pulled after academics find breach of doctor details possible

By political reporter [Stephanie Anderson](#)
Updated 29 Sep 2016, 2:51pm

The Health Department has removed Medicare data from its website amid an investigation into whether personal information has been compromised.

Australian Privacy Commissioner Timothy Pilgrim has launched an investigation after academics found it was possible to figure out some service provider ID numbers in the Medicare Benefits Schedule and Pharmaceutical Benefits Schedule datasets, published on August 1.

The University of Melbourne academics said they notified the department of the issue on September



PHOTO: The Health Department says no patient information has been compromised. (ABC News)



Privacy Amendment (Re-identification Offence) Bill 2016

*One day before agreed
announcement day...*

Attorney-General memo:

Intention to amend *Privacy Act 1988*

- Legislative instrument permitting retroactive prosecution

The Bill

- Criminalises **reID** of Commonwealth data
- Up to **2 year jail** term
- **Reverses burden** of proof
- **Retrospective** legislation (see memo above)
- Exemptions for AG or Dept-sanctioned investigations
- Apparently* might not apply to all* academics

The screenshot shows the ZDNet homepage with a search bar and navigation links for Videos, Executive Guides, Mobility, Security, NBN, CXO, Hardware, and Newsletters. A red banner at the top reads "MUST READ: Australian National University breached with 19 years of data accessed". The main article title is "Brandis swings his golden hammer, misses mark".

Brandis swings his golden hammer, misses mark

Criminalising the re-identification of de-identified government data will hinder legitimate researchers and do nothing to improve citizens' privacy.



By Stilgherrian for The Full Tilt | September 29, 2016 -- 06:52 GMT (16:52 AEST) | Topic: Security

Life of a ReID Bill

- 12 Oct 2016 – Intro to the Senate
- 10 Nov 2016 – Ref to Senate Legal Committee
- 16 Dec 2016 – Consultation Period Ended
- 7 Feb 2017 – Senate Committee Report
- . . .
- 6 Jun 2019 – *Zombie Bill!*



Overwhelmingly critical response

Law Council of Australia

- “*The reverse onus provisions should be removed ...*”
- “*definition of ‘de-identified’ ... not always clear cut*”

Australian Bankers Associations

- “*re-identification ... might accidentally occur, without the bank intending*”

14/15 submissions critical

Final Senate Committee report:

“*The committee notes the concerns ... However...the bill provides a necessary and proportionate response*”

ZDNet

MUST READ: Australian National University breached with 19 years of data accessed

Brandis re-identification law proposal slammed

While Health Minister Sussan Ley has apologised for the breach in de-identified medical data, Australia's Attorney-General has come under fire for 'rushing' through legislation.



By Asha Barbaschow | September 30, 2016 -- 04:26 GMT
(14:26 AEST) | Topic: Security



Advertisement

NATIONAL PUBLIC SERVICE

Fears that patients' personal medical information has been leaked in Medicare data breach

By Rania Spooner and Noel Towell

Updated September 29, 2016 — 4:37pm, first published at 11:05am



There are fears patients' sensitive medical information could have been made public in a Medicare data breach by the health department.

Health Minister Sussan Ley insists the data, which was loaded onto the internet, does not identify patients.

“Health Minister Sussan Ley insists the data, which was loaded onto the internet, **does not identify patients.”**

TODAY'S TOP STORIES
Profits before people: ANZ defies government, as CBA and NAB pass on cut
2 hours ago

But, we're abound with health data

Who has been hit by the injury jinx a your club?

ANDY OTTEN (ADELAIDE)

Drafted: 2007

Games: 79 (average 11 per year)

Drafted with pick 27 in the 2007 national draft, Otten has had an injury-cursed career after a breakout season in 2009. In a cruel twist of fate, it was at the end of that year that he suffered his first serious knee injury and required a knee reconstruction. Since then he has had a number of knee worries and before again rupturing his ACL in August last year.

He's not the only Crow who's a regular in the rehab group, with young gun Brad Crouch suffering a broken leg, broken foot and Achilles worries in his first three seasons.

Note: average games does not include 2015 season.

Natalie Terese Abberfield was born in 1969.¹ She is the daughter of Kenneth Beresford Abberfield and Melissa Judith Dutton.¹ She married Barnaby Joyce in 1993.¹

Children of Natalie Terese Abberfield and Barnaby Joyce

1. Bridgette Maree Joyce¹ b. 1996
2. Julia Frances Joyce¹ b. 1998
3. Caroline Wilga Joyce¹ b. 2000
4. Odette Honora Joyce¹ b. 2002

The Sydney Morning Herald

YOUR SHORTLIST IS EMPTY
Add stories using the  button

Tanya's Plibersek juggle of love

By ROSIE SQUIRES, The Sunday Telegraph
October 10, 2010 12:12am



AT only 10 days old, Labor Minister Tanya Plibersek's newborn carries the weight of a nation.

Soon afterwards, Bernardi was diagnosed with tuberculosis and put in isolation, where he was asked to list all the people he had been in recent contact with.

"I didn't have the nerve to tell them [about] John Howard," Bernardi admits in the latest episode of ABC's *Kitchen Cabinet* to air on Wednesday night.

"I thought, let him go to the election and then we'll deal with it."

But when Howard was later hospitalised with lung problems, Bernardi was convinced he'd "done John Howard in".



Searching for Vanessa

17,310 women share her birthyear

59 also had children born 2006, 2011 in Australia

23 also based in Victoria

0 with child DOB with perturbations

→ **Not** in dataset

```
1 v {"$GENDER": "F", "YOB": 1975,          { $and: [
2 v           { $and: [
3 v             {"MBS": { $elemMatch: { "ITEM": { $in: ["00200", "00201", "00204", "00205", "00207",
4 v               {"DOS": { "$gte": "2000-01-01 00:00:00", "$lte": "2000-01-01 00:00:00" }
5 v             }
6 v           }
7 v         },
8 v         {"MBS": { $elemMatch: { "ITEM": { $in: ["00200", "00201", "00204", "00205", "00207",
9 v               {"DOS": { "$gte": "2000-01-01 00:00:00", "$lte": "2000-01-01 00:00:00" }
10 v             }
11 v           }
12 v         },
13 v         {"MBS": { $elemMatch: { "ITEM": { $in: ["00200", "00201", "00204", "00205", "00207",
14 v               {"DOS": { "$gte": "2000-01-01 00:00:00", "$lte": "2000-01-01 00:00:00" }
15 v             }
16 v           }
17 v         }
18 v       }
19 v     } }
```

By default, valid **ObjectID** and **ISODATE** (YYYY-MM-DD HH:mm:ss) strings are being converted into MongoDB **ISODate** strings. You can change this behaviour

Execute

nt - patients x count - patients x count - patients x count - patients x

Tree ▾

: 1

Anyone could do this!!



Mothers unique in MBS-PBS-10%

Child birth years $[b, b + 4]$

Mum birth years $[m, m + 5]$

... for suppressed $m \gg b$

Similarly for small m

Year of birth (mother\baby)	b	b+1	b+2	b+3	b+4
m	0	0	0	1	1
$m+1$	0	0	1 ³	1	0
			Dec b+2 or Jan b+3	Dec b+2 or Jan b+3	
$m+2$	0	1	1	0	1
		July/Aug	July		
$m+3$	1	0	0	0	0
$m+4$	1	1	1	0	0



DeID as property of datasets... (compare to DP)

- **Unique in the Crowd: The privacy bounds of human mobility** - *Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen & Vincent D. Blondel*
- **Not So Unique in the Crowd: a Simple and Effective Algorithm for Anonymizing Location Data** - *Yi Song, Daniel Dahlmeier & Stephane Bressan*
- **Unique in the shopping mall: On the reidentifiability of credit card metadata** - *Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh, Alex "Sandy" Pentland*
- **Big Data and Innovation, Setting the Record Straight: De-identification Does Work** - *Ann Cavoukian & Daniel Castro*
- **No silver bullet: De-identification still doesn't work** - *Arvind Narayanan & Edward W. Felten*
 - ✓ *ReID models overly narrow by assuming only access to public data*
 - ✓ *ReID does not require specialist skills*
 - ✓ *Quantification of reID probabilities at best lower bounds/fundamentally meaningless*



It's only a sample; bah humbug “*confidence*”!!

DHS whole-of-population statistics on MBS billing rates

- aggregated: 10yr age ranges, billing state, billing month
- 27% of codes are uniquely reported
- E.g. aortic valve replacement, Former PM in Brisbane, Aug 2011 (in pop, not sample)



Reidentifications

Wiki/news articles on 18 mums with 2+ births

- 13 had no matches, including
Gillian Triggs – Former Human Rights Commis.
Natasha Stott-Despoja – Former Senator
Cathy Freeman – Olympic Athlete
Tanya Plibersek – Current MP
- 2 rejected due to inconsistent information
- 3 returned a unique match

Other risks – only partially assessed

Fingerprinting by billing amounts
Melbourne Pharmaceutical Datathon with zipcodes

25 more queries

- Professional footballers with injuries
- Politicians with reported unusual surgeries
- 4 more unique matches
- One is an AFL team captain matching medical history, birth year, and interstate movements
- One has likely been reported as the oldest person in a state to have received a surgery, as confirmed by whole-of-population data



A release born out of an open-data-first environment

IT IT SECURITY FINANCE IT TELCO BENCHMARK AWARDS

[Twitter](#) [Facebook](#) [LinkedIn](#) [RSS](#) [Login](#) [Subscribe](#) [Search](#)

Govt embraces Five Safes in new data sharing guidance



By Justin Hendry
Mar 19 2019
11:17AM



The Commissioner's view is that the processes or steps for achieving re-identification are so extensive, and the risk of identification for any given patient so low, that the patients in the dataset are not reasonably identifiable for the purposes of the Privacy Act.

On the information before him, the Commissioner's view is that the published dataset does not contain the personal information of patients, as defined in the Privacy Act, as patients cannot be reasonably identified.



Where to from here?

Data can hold huge potential societal benefit

Provide access to people who need it, whilst protecting privacy

Directions

- List what data is held – transparency in who holds what data
- Secure research environments – e.g. ONS Virtual Microdata Laboratory
- Crypto, SMC, private linkage
- *More* reID for assessing privacy risk
- **Differential privacy with many advantages** (much over the next 2 days)



Partners in it's-not-a-crime



Dr Chris Culnane



Associate Prof Vanessa Teague

We also variously like
Diff. privacy (TransportNSW)
Applied crypto (Swiss sVote)
Bayesian record linkage (ABS)

...

Visitors welcome!!