inspiring success

Australian Computer Society Inc. (ACT)
ARBN 160 325 931

Level 27, Tower 1,
100 Barangaroo Avenue
Sydney, NSW 2000
T 02 9299 3666

Privacy Preserving Data Sharing Frameworks

Report on July 2019 Directed Ideation #2 Series
Version 1.0
9th August 2019

Editor: Ian Oppermann

Table of Contents

1. BACKGROUND 4
1.1 The Problem 4
1.2 Personal Information 4
1.3 A Modified "Five Safes" Framework 4
1.4 A Personal Information Factor 5
1.5 Framework for Considering PIF 6
1.6 Data Sets Used 7
1.7 Directed Ideation Series - How it Worked 8
2. STARTING POINT 10
2.1 Personal Information Factor – A Starting Point 10
2.2 Mutual Information as a Measure of Utility 10
2.3 Dealing with Trajectories – A Starting Point 11
3 THE RELATIONSHIP BETWEEN MUTUAL INFORMATION AND PIF 13
3.1 Assessing Features in a Dataset Based on Feature Information Gain (FIG) 13
3.2 Feature Dependence 14
3.3 Matrix of Mutual Information 15
3.4 Implementation 17
4 DEALING WITH TRAJECTORIES 19
4.1 Trajectory Flattening Techniques 19
4.2 Depth Information Gain 19
5. PROTECTING DATA THROUGH PERTURBATION 23
5.1 Perturbation through Random Noise is Different 23
5.2 A Differential Privacy Approach 23
6. CONCLUSIONS 27
7. THANKS 28
APPENDIX A – February 2019 Event PIF Model 29
APPENDIX B – SAMPLE DATASETS 35

Executive Summary

The paper reports on a three-week activity in July 2019 which explored aspects of the 2018 ACS Technical Whitepapers on Privacy preserving Data Sharing. The three-week activity further developed frameworks explored in a similar event in February 2019. Both events attempted to develop a Personal Information Factor (PIF) and developing Data Safety factors based on the description in the technical whitepapers. The PIF and frameworks were evaluated on open datasets and synthetic datasets.

Conclusion 1: The use case for data strongly influences the risk framework for data safety and the methods (aggregation, generalisation, obfuscation, perturbation) appropriate for increasing data safety.

Conclusion 2: Development of a meaningful measure for Personal Information Factor (PIF) is feasible. Information theoretic metrics such as PIF show promise as a way to measure the privacy risk of unit record data for aggregated, generalised or obfuscated data, and can be enhanced to cover perturbed data. Additional work would need to be done to relate the privacy risk metric to the legal definition of privacy, and the assumed attacker model.

Conclusion 3: Development of a meaningful measure of relative Utility is feasible for datasets which have been protected through aggregation, generalisation, obfuscation and perturbation. Information Theoretic metrics based on Mutual Information (between original and protected datasets) show promise.

Conclusion 4: Dealing with "trajectories" is a critical problem for the release and use of datasets. Development of means to deal with datasets linked to form a trajectory are possible. The methods explored shows some promise, however the complexity of the approaches may limit real-world implementation.

Conclusion 5: Understanding the relationship between different features in a dataset helps to identify those features which pose the highest risk of reidentification and those which have the greatest impact on utility after protection methods are applied.

1. BACKGROUND

1.1 The Problem

Future Smart Services for homes, factories, cities, and governments rely on sharing of large volumes of often personal data between individuals and organisations, or between individuals and governments. The benefit is the ability to create locally optimised or individually personalised services based on personal preference, as well as an understanding of the wider network of users and providers. Despite these benefits, data sharing remains a challenge for several privacy-based reasons:

- There is currently no way to unambiguously determine if there is personal information in aggregated data. De-identification and aggregation are common approaches used to reduce the level of personal information in a dataset when linking or releasing. Different deidentification approaches and different levels of aggregation are used by organisations depending on a perceived value of an associated risk. The implications of this are profound when thinking of the use cases which come in and out of scope depending on the level of aggregation used.
- Concerns raised by Privacy advocates as the capability of data analytics increases. When the number of datasets used to create a service or address a policy challenge increases to hundreds or thousands, the complexity of the problem may rapidly exceed the ability of human judgement to determine if the combined data (or the insights generated from them) contain personal information.
- Context: A linked dataset may have low information content for one observer, and high for another who brings with them their unique knowledge and history. What in a limited context may be the identification of any single individual ("any" anyone), may become identifiable with an actual individual (an actual "someone").

1.2 Personal Information

Personal data covers a very wide field and is described differently in different parts of the world. For example, in NSW:

> *"… personal information means information or an opinion (including information or an opinion forming part of a database and whether or not recorded in a material form) about an individual whose identity is apparent or can reasonably be ascertained from the information or opinion".*

The legal tests for personal information generally relate to the situation where an individual identity can *"..reasonably be ascertained"*. The definition is very broad and in principle covers any information that relates to an identifiable, living individual for 30 years after their death.

1.3 A Modified "Five Safes" Framework

In September 2017, the Australian Computer Society (ACS) released a technical whitepaper which explored the challenges of data sharing[1]. The paper highlighted that one fundamental challenge for the creation of smart services is addressing the question of whether a set of datasets contains personal information. Determining the answer to this question is a major challenge as the act of combining datasets creates information. The paper further proposed a modified version of the "Five Safes" framework[2] for data sharing which attempts to quantify different thresholds for "Safe". In November

[1] See ACS website, available online https://www.acs.org.au/content/dam/acs/acs-publications/ACS_Data-Sharing-Frameworks_FINAL_FA_SINGLE_LR.pdf (accessed 7th March 2019)

[2] T. Desai, F. Ritchie, R. Welpton, *"Five Safes: designing data access for research"*, October 2016, "http://www.nss.gov.au/nss/home.NSF/533222ebfd5ac03aca25711000044c9e/b691218a6fd3e55fca257af700076681/$FILE/The%20Five%20Safes%20Framework.%20ABS.pdf (accessed 7th March 2019)

2018, the ACS released a second technical whitepaper on Privacy Preserving Frameworks[3] which evolved the concepts introduced in the first paper.

The whitepapers introduced several conceptual frameworks for practical data sharing including an adapted version of the "Five Safes" framework. Several organisations around the world including the Australian Bureau of Statistics use the Five Safes framework to help make decisions about effective use of data which is confidential or sensitive (including; because of the presence of personal information).

1.4 A Personal Information Factor

The ACS Technical whitepapers explored a hypothetical parameter, the "Personal Information Factor" (PIF) which was a measure of the personal information in a linked, deidentified dataset or in the outputs of analysis.

A PIF of 1 means sufficient personal information exists to identify an individual: the total personal information (PI) is personally identifiable (PII). A value of 0 means there is no personal information. It is important to note that the PIF described is not a technique for anonymisation: rather, it is a heuristic measure of potential risk of reidentification.

The PIF for both data and outputs is described based on:

- A measure of the information content of the dataset used to conduct analysis or the output of the analysis (the simplest analysis may be sharing of data);
- The uniqueness of the most unique individual (group) in the dataset or output;
- Additional information required by the observer to be able to identify an individual from the data or outputs.

Figure 1 shows how PIF may be considered when project data or outputs have been released into a range of different environments, from those which can be controlled to the broadest possible environment, open data. With open data, there is no control over who accesses the dataset or the analysis outputs, and which additional datasets can be combined with the outputs.

[3] See ACS website, available online https://www.acs.org.au/content/dam/acs/acs-publications/Privacy%20in%20Data%20Sharing%20-%20final%20version.pdf (accessed 7th March 2019)

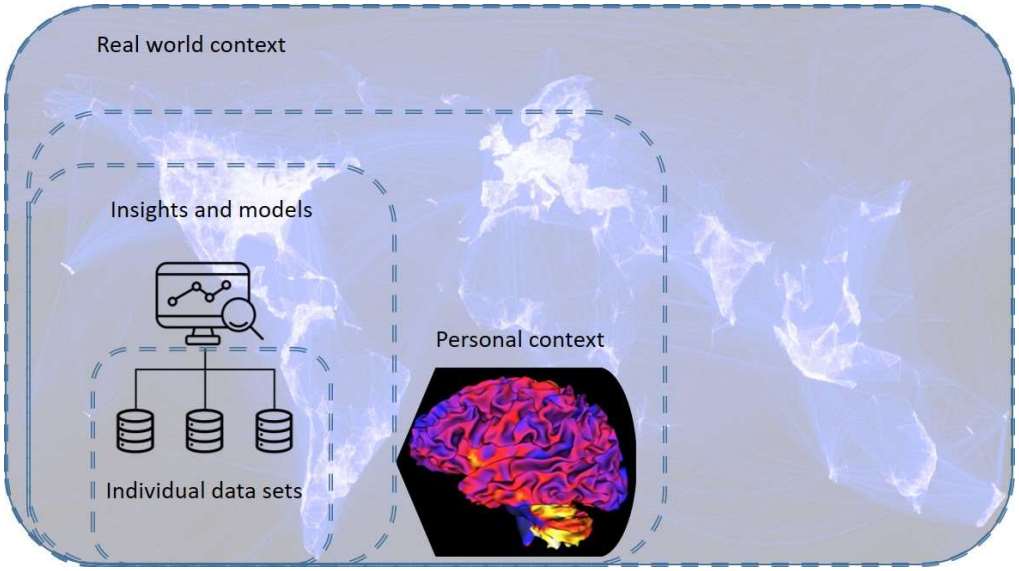


Figure 1. Real world context for evaluating PIF

1.5 Framework for Considering PIF

An attacker or threat model is adopted as summarised in Figure 2. This is motivated by the fact that with every dataset released, there is an increase in the information available about a person. However, not every reidentification event is of the same severity. For example, learning that person X's previous year's income was between $50k to $150k reveals less personal information than learning the exact figure. So instead of focusing on the individual's reidentification risk, the framework computes the potential data gain for each field (individual value, for example, one individual's salary) in the dataset.

The framework then allows the user to:

- reason about risks on a per-feature (per-column in the table) basis,
- find risks of particular individuals (rows in the table),
- identify comparatively high-risk individuals,
- inform anonymisation efforts on where to focus, and
- compare different anonymisation strategies.

Formalising the threat model by using a concept from cryptography: the attacker model. An attacker is a person who has access to the dataset and to some additional information about a particular individual. They wish to locate that individual in the dataset to learn more information about them. Knowing the true strength of a potential attacker is difficult as it is hard to correctly quantify the auxiliary information available to the attacker.

In the absence of better information, an attacker can be modelled as very powerful: they know every feature of a person aside from the one they are attempting to find. Nonetheless, models that are less strong are also possible: we could assume that they know some but not all features, or that they are not fully certain in the information that they have. Using this approach, the quantified concepts of Cell Information Gain (CIG) and Row Information Gain (RIG) were developed in the February 2019 event (see Appendix A for further information).

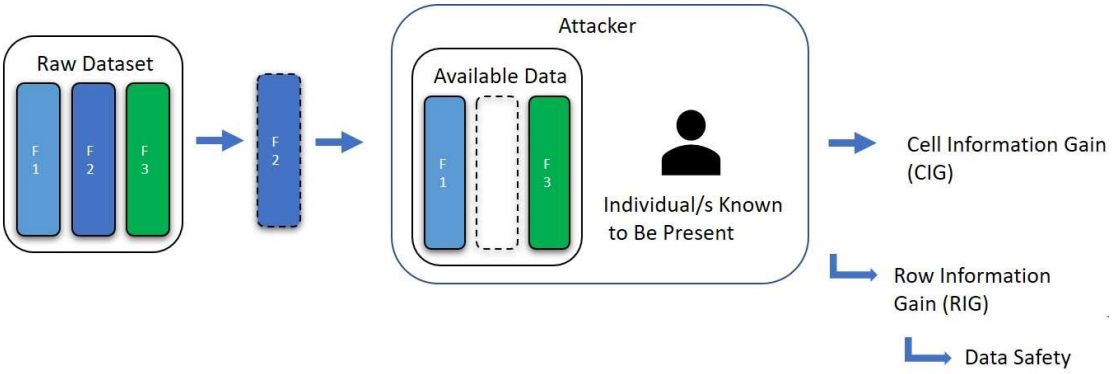


Figure 2. Overview Approach

1.6 Data Sets Used

Three core datasets were used in the competition, a number of synthetic datasets and open datasets were used (see Appendix for more details and samples):

- Dataset 1 – Inmate Admissions (United States open dataset)
  Inmate admissions with attributes (race, gender, legal status, top charge). Unit record level with unique identifier of inmates. An inmate can have multiple charges, status, admission time, and discharged time. 301,748 rows and 7 columns, with 148k unique inmate ID's.

- Dataset 2 – Open Parking and Camera Violations (United States open dataset)
  This dataset contains Open Parking and Camera Violations issued by the City of New York Record level on vehicle plate number with violation, and issue date. One vehicle plate can have multiple violations over time. 39.4m rows and 19 columns

- Dataset 3 – Air BNB Sydney Listings (commercial open dataset)
  Publicly available information pooled by Inside Airbnb, with host ID, name, property listings, price, coordinates, text description.

- Dataset 4 - NYC Green Taxi Trip Data (United States open dataset)
  The green taxi trip records include fields pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.

- Dataset 5 - ATO Taxation Individual Statistics (Australian open datasets)
  Aggregated individual taxation statistics by industry consisting financial year 2013-14, 2014-15, 2015-16, and 2016-17 (four separate datasets combined). Included are description of industry, amount of tax, taxable income, Medicare levy and superannuation.

- Dataset 6 - Synthetic NAPLAN Test Result Data (Synthetic dataset)
  Randomly generated unit record level of student performance on the NAPLAN test. Each record has a student's name, country of birth, year level, one parent's occupation group, School ID, and the test results in the form of bands. The randomly generated test result consists of reading, spelling, grammar and punctuation, writing, and numerical literacy. Data is randomly generated however adheres to the major statistical properties of the original dataset.

- Dataset 7 - Synthetic Hospital Admissions Data (Synthetic dataset)
  Randomly generated dataset with fields including personal information (name, address, DOB, occupation) as well as medical diagnosis from GBD [4] (Global Burden of Disease). The prevalence distribution of the medical conditions by age group and gender in Australia can be accessed using tools provided by GHDx[5]. Unit record level detail (synthetic) patients admitted to the hospital with diagnosis details, date of birth, gender, occupation, and address. Each individual synthetic patient has a trajectory of different visit time and diagnosis.

- Dataset 8 - Synthetic NSW People Matter Employee Survey (PMES) (Synthetic dataset)
  Randomly generated dataset with fields including demographic attributes of the survey respondents (education level, age group, disability status, employment status, gender, LGBTI status, and ethnical diversity) along with the Likert scale responses to the survey questions.

- Dataset 9 – Synthetic NSW Workforce Profile Data (Synthetic dataset)
  Randomly generated dataset with fields including personal information (DOB, gender, country of birth, minority group status, highest education level, and disability status. Each individual synthetic government employee has a trajectory of changes in remuneration, legislation code, salary band, and standard weekly full-time hours over three years.

1.7 Directed Ideation Series - How it Worked

The Directed Ideation is intended to bring to life some of the major aspects of the Data Sharing frameworks described in the 2018 ACS Technical Whitepaper and advanced in the February 2019 Directed Ideation. Over a three-week period in July 2019, teams competed to develop risk frameworks based on the PIF, evaluate how "safe" each dataset was, and to create "safer" versions of each dataset. The teams were also encouraged to improve the way PIF and Utility are determined and to explore improved ways of dealing with datasets with trajectories.
The summary of tasks for teams is shown in Figure 3.

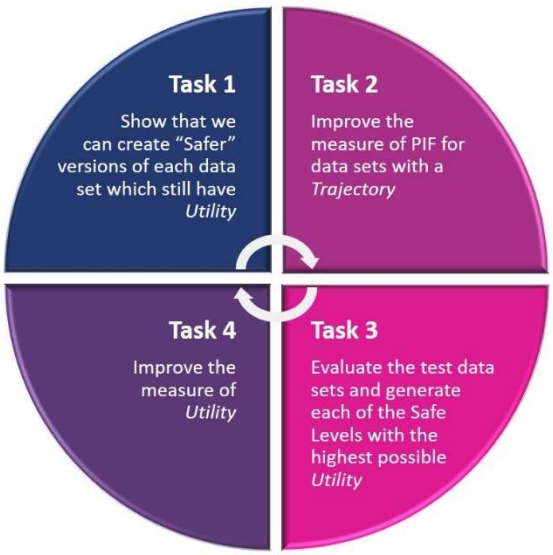


Figure 3. Challenge Tasks

[4] See https://www.who.int/healthinfo/global_burden_disease/about/en/ (accessed July 2019).
[5] See http://ghdx.healthdata.org/gbd-results-tool (accessed July 2019).

Run as a competition, a total of 16 people worked in teams on a number of tasks each week. At the end of the first week, a down-selection process took place (see Figure 4) with teams growing in size but reducing in number. In the final round of the competition, three approaches were presented from 3 colourfully named teams:

- Team 1 – Good Fighters (GoFi)
- Team 2 – Baysically Measure Zero (BaMeZe)
- Team 3 – Privately Concerned (PriCo)

Each team approach was the result of several approaches combined and refined through direction from judges at the end of the previous round.

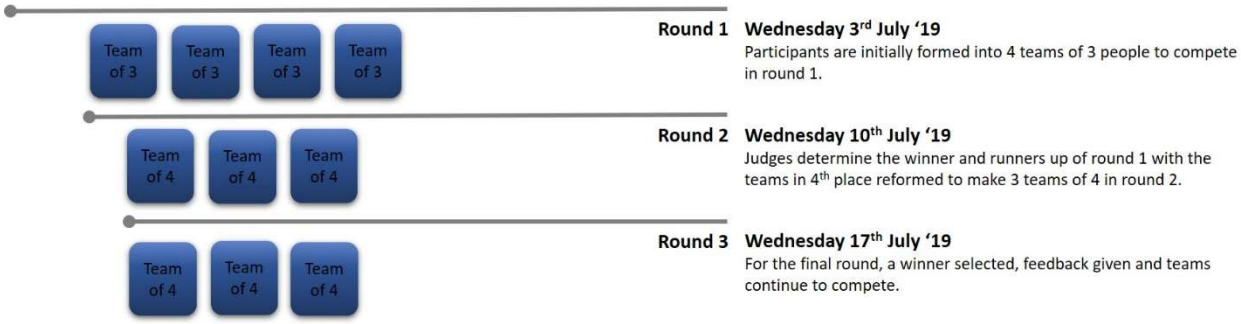


Figure 4. Competition Rounds

2. STARTING POINT

2.1 Personal Information Factor – A Starting Point

Development of the Personal Information Factor was a goal of the Directed Ideation and it was expected to evolve during the course of the event.

The starting point was based on the concepts of information gain developed by the teams in the February 2019 event (see Appendix A for more detail).

Cell Information Gain (CIG) is used to quantify the reidentification risk for each piece of personal information. Every cell belongs to a row, and every row represents information about a person. We imagine that an attacker is attempting to reidentify a person to find the value of the cell whose CIG we are determining. We assume the attacker knows every feature of this person except this one cell. Its CIG is then defined as the KL-divergence of the attacker's prior and posterior beliefs for the true value of that cell.

The CIG is calculated (in bits) as

$$\begin{aligned} D_{\mathrm{KL}}(P \parallel Q) &= -\sum_{x \in \mathcal{X}} p(x) \log q(x) + \sum_{x \in \mathcal{X}} p(x) \log p(x) \\ &= \mathrm{H}(P, Q) - \mathrm{H}(P) \end{aligned}$$

By summing the CIG for every row, or individual, of the table to obtain a *Row Information Gain* (RIG). It measures how susceptible a particular individual is to having their information revealed through reidentification in the dataset.

Similarly, by summing the CIG for every row to find the *Feature Information Gain* (FIG) for that feature. The FIG is a measure, in bits, of the reidentification risk of a feature. It can help us to identify the features that are the highest risk to include in a dataset.

2.2 Mutual Information as a Measure of Utility

In probability theory and information theory, the mutual information (MI) of two random variables is a measure of the mutual dependence between the two variables. More specifically, it quantifies the "amount of information" (in bits) obtained about one random variable through observing the other random variable. The concept of mutual information is intricately linked to that of entropy of a random variable, a fundamental notion in information theory that quantifies the expected "amount of information" held in a random variable.

Not limited to linear dependence like the correlation coefficient, MI is more general and determines how similar the joint distribution of the pair ***(X,Y)*** is to the product of the marginal distributions of ***X*** and ***Y***. MI is the expected value of the pointwise mutual information (PMI) and is known as information gain (or loss).

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p_{(X,Y)}(x,y) \log \left(\frac{p_{(X,Y)}(x,y)}{p_{(X)}(x) p_{(Y)}(y)} \right)$$

In this Directed Ideation, MI was normalised to values between 0 and 1 by dividing it against the mutual information of the original data itself ***I(X; X)*** to produce a measure of utility ***μ*** for a "Safer" version of a dataset compared to the original dataset.

$$\mu = \frac{I(X;Y)}{I(X;X)}, I(X;X) \neq 0$$

A utility of 1 implies no information loss and ideal utility. A value of 0 implies complete information loss and no utility in the resultant dataset. Figure 5 shows a simple example of Utility declining as a feature "age" is aggregated in a dataset.

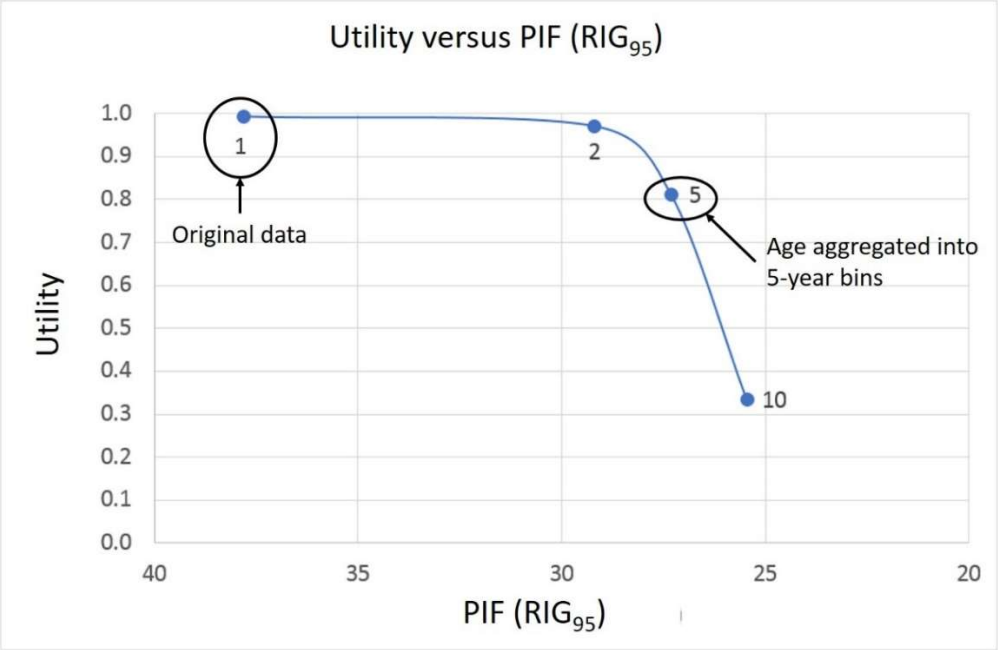


Figure 5. Example of decrease in Utility and PIF as "age" is aggregated into broader bins

2.3 Dealing with Trajectories – A Starting Point

Development of the means to accommodate trajectories (linked rows) was a further goal of the Directed Ideation and it was expected to evolve during the course of the Directed Ideation. The starting point however was based on subsequence decomposition which considered both continuous and non-continuous sub-sequences of all features which could be used to create a trajectory. For example, with the hospital admissions dataset, a trajectory for each patient could be constructed based on time of visit, hospital venue or reason for admission. Each of these features, or combinations of them could be used to identify a unique trajectory for the individual. If the full sequence of visits was not known, but knowledge of a unique ordering of visits existed (even without knowledge of visits between stages of this unique ordering), then it is possible to identify a unique trajectory.

The overall approach is outlined below:

- Every individual has a linked set of rows which forms a sequence. Each of these linked rows has a number of features which could form a trajectory in isolation or as groups.
- Find all possible (continuous and non-continuous) sub-sequences from the main sequence. Group them into 1-step, 2-step, …. N-step sub-sequences.
- For each possible sub-sequence in the dataset and each number of steps, determine the number of individuals with this sub-sequence.
- The sub-sequences with the greatest privacy risk are those associated with only one individual

- Determine the maximum allowed number of steps that does not contain only one individual.
- Repeat this process for all features which can form a trajectory.

A worked example for the dataset 1 (Inmate admissions) is shown in Figure 6. The prison sites "DE", "CS" refer to individual venues. The full sequences are provided for each individual inmate, and for non-trivial sequences, it can readily be seen that the number of length n sequences can be calculated from the number of unique venue transitions (continuous and non-continuous) as shown in Figure 7.

id	Admission Sequence
19	['DE', 'DE']
20	['DE', 'DE', 'CS', 'CS', 'DE', 'DE', 'CS', 'DE', 'DE']
21	['CS']
22	['DE']
23	['DE', 'DE', 'DE', 'DE', 'DE', 'DE', 'DE']
24	['SSR', 'DE']
25	['CSP', 'DE', 'DE', 'CSP', 'DPV', 'DPV', 'SCO', 'SSR', 'DEP']
26	['DE', 'DPV', 'DE']
27	['DE']

Figure 6. Sample of Inmate Admissions dataset with trajectory based on prison site

No. of subsequences	No. of individuals with subsequence									
No. of steps	1	2	3	4	5	6	7	8	9	10
1										
2		1		1	1	2		1		
3	39	26	24	12	18	9	10	7	5	3
4	410	231	154	106	68	53	61	53	47	32
5	2094	798	419	339	184	153	109	101	88	61
6	5160	1452	680	398	307	210	143	102	67	50
7	8266	1760	712	356	218	123	108	99	77	56
8	9632	1503	601	326	244	200	160	150	87	75
9	9099	1511	740	478	298	165	111	91	56	31
10	8456	1660	634	220	82	37	12	4	1	
11	7167	817	127	26	15	25	48	71	103	101
12	4517	673	653	628	539	426	278	192	157	73
13	4449	1309	435	184	89	35	20	9	7	3
14	2262	198	33	3	2	1				
15	368	5	1							

No sub-sequence of length 1 or length 2 is associated with only 1 individual

Each of these 3-step sub-sequences is associated with only 1 individual

Each of these length-3 sub-sequences is associated with only 2 individuals

Figure 7. Worked example calculating number of sub-sequences of differing lengths

This approach highlights the challenge of datasets with trajectory characteristics. In the Inmate Admissions example above, the length of known (continuous or non-continuous) subsequence is only length 3 before 39 unique records can be identified. At length 4, this rises to 410 unique records.

3 THE RELATIONSHIP BETWEEN MUTUAL INFORMATION AND PIF

3.1 Assessing Features in a Dataset Based on Feature Information Gain (FIG)

The concept of Cell Information Gain is based on a KL-Divergence measure of information gained (in bits) of an attacker who gains knowledge of an actual cell value compared to the prior believed value of that cell. This concept allows us to consider individual features from the perspective of risk of reidentification. Figure 8 shows the minimum, maximum, average and quartile band values for features in dataset 6 (synthetic NAPLAN test results data). This figure shows that school ID and data of birth (DOB) are high risk features from a reidentification perspective. Country of birth however is a relatively low risk factor for most individuals in the dataset population except for a small number for whom it is a high-risk factor. This highlights the real-world challenge of outliers in a dataset being susceptible to reidentification. Gender is seen to be low risk for the entire population indicating a balance of genders in the dataset population.

	Min	Q1	Avg	Med	Q3	Max
SchoolID	7.38	8.96	9.43	9.96	9.96	9.96
Surname	5.57	7.64	8.53	8.96	9.96	9.96
First_Name	5.06	6.79	7.82	7.64	8.96	9.96
Gender	0.98	0.98	1.00	0.98	1.02	1.02
DOB	7.64	8.96	9.36	9.96	9.96	9.96
Year_Level	1.89	1.89	2.00	1.90	2.08	2.14
Student_Country_of_birth	0.21	0.21	1.23	0.21	0.21	9.96
Parent1_Occup_Group	1.97	1.97	2.53	2.56	2.69	3.13
readband	2.00	2.22	2.89	2.69	3.79	6.64
splband	2.21	2.44	2.90	2.52	3.61	6.96
grpnband	2.13	2.44	2.88	2.62	3.50	6.79
writband	1.87	1.87	2.69	2.00	3.25	7.96
numband	2.23	2.50	2.97	2.71	3.21	7.38

Figure 8. FIG bands for features of dataset 6 (Synthetic NAPLAN Test Result Data)

As the numerical valued features in the dataset are aggregated, the values of the FIG bands change as shown in Figure 9. The aggregation performed in this example considers every feature to be independent. As features are aggregated, the FIG change in almost all bands. Nonetheless, the challenge of outliers remains indicating that further aggregation is required.

	Min	Q1	Avg	Med	Q3	Max
SchoolID	0.82	0.82	4.24	0.82	8.96	8.96
Surname	1.53	1.53	6.04	7.16	9.96	9.96
First_Name	4.08	6.64	7.66	7.64	8.96	9.96
Gender	0.98	0.98	1.00	0.98	1.02	1.02
DOB	0.67	0.67	3.71	0.67	8.96	8.96
Year_Level	1.89	1.89	2.00	1.90	2.08	2.14
Student_Country_of_birth	0.18	0.18	1.04	0.18	0.18	9.96
Parent1_Occup_Group	1.97	1.97	2.53	2.56	2.69	3.13
readband	2.00	2.22	2.89	2.69	3.79	6.64
splband	2.21	2.44	2.90	2.52	3.61	6.96
grpnband	2.13	2.44	2.88	2.62	3.50	6.79
writband	1.87	1.87	2.69	2.00	3.25	7.96
numband	2.23	2.50	2.97	2.71	3.21	7.38

Figure 9. FIG bands for aggregated features of dataset 6 (Synthetic NAPLAN Test Result Data)

3.2 Feature Dependence

The concept of *Cell Information Gain* is based on a KL-Divergence measure of information gained (see Appendix A) when an attacker learns the true value of a cell as opposed to the prior assumed value. If the learned value and prior values are the same, there is no information gain (0 bits).

Throughout the various phases of the exercise, the ability to infer information between features has been largely ignored. The introduction of the concept of "mutual information" allowed an exploration of feature dependence and gave insights into which features represented the highest risk of reidentification. The significance of feature dependence is that it impacts the incremental level of information gained once the true value of a feature is learned.

Figure 10 shows the mutual information between all pairs of features in dataset 6 (Synthetic NAPLAN Test Result Data). A high value refers to a high level of mutual information. In the original dataset (LHS), the diagonal contains high MI values for most features indicating a balanced (not highly skewed) distribution for the feature. A low level on the diagonal indicates a distribution with outliers as seen in the feature "Student_County_of_birth". Similarly, features "readband" and "writband" show values significantly less than 1. Off the diagonal, there are small but non-zero values between features "DOB" and "SchoolID", and between features "DOB" and "Surname" indicating some mutual information between features or, feature dependence within this dataset.

In the aggregated dataset (RHS), the MI has again been calculated between all feature pairs. The off-diagonal values have been reduced to zero removing the feature dependence. On the diagonal, the values for "Student_County_of_birth" and "Surname" have increased implying a less skewed distribution for the feature. However, the MI for features such as "SchoolID" has decreased. The implication is that, for this particular aggregation technique, the dependence between features has been removed, but the approach has made the distribution more skewed implying introduction of more outliers. Not all protection-through-aggregation techniques are the same.

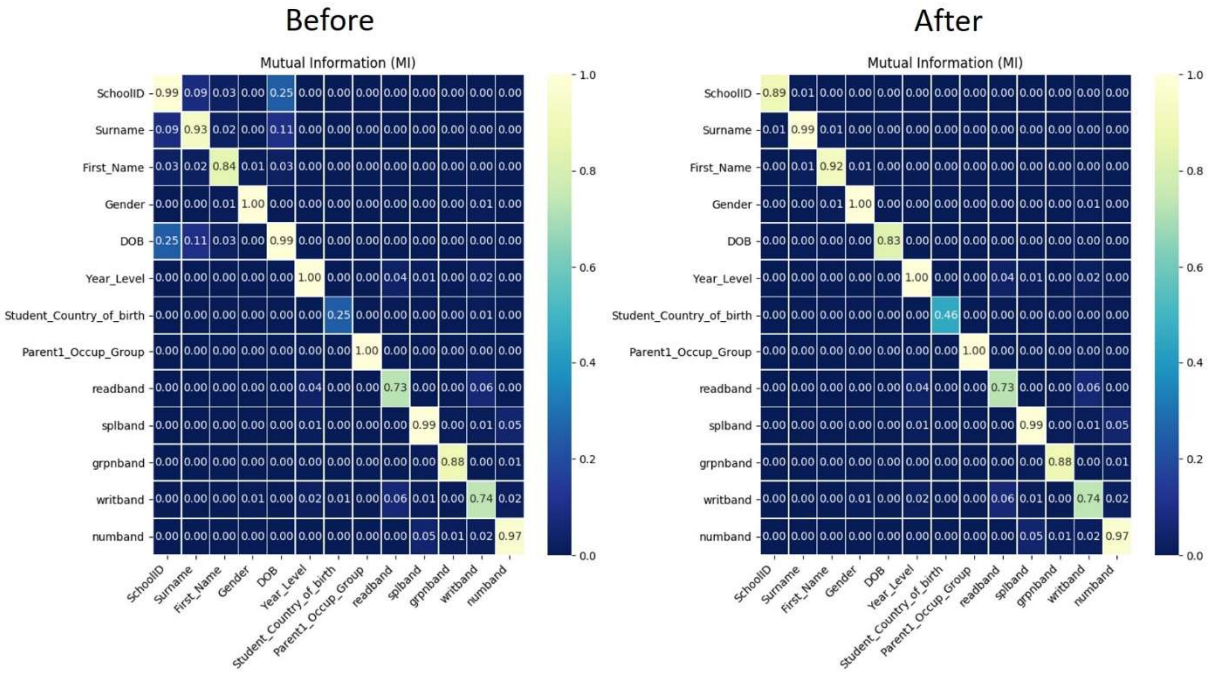


Figure 10. Mutual Information between features before and after aggregation

3.3 Matrix of Mutual Information

Understanding the relationship between features in a dataset provides insights as to how to create "safer" versions of the dataset.

As an example, the relationship between features in dataset 8 (NSW Public Sector Workforce Synthetic dataset) "age", "salary" and "years in job" can be examined as they are independently aggregated. The focus for aggregation are features with ordinal values and effectively treats each feature as an independent dataset. Figure 11 shows the change in PIF (RIG_{95}) for each single-feature dataset versus the loss in mutual information between original and aggregated dataset. From this figure, it could be concluded that there the significant reduction in PIF from aggregating "salary" makes it a more obvious target for protection through aggregation compared to "age" and "years in job".

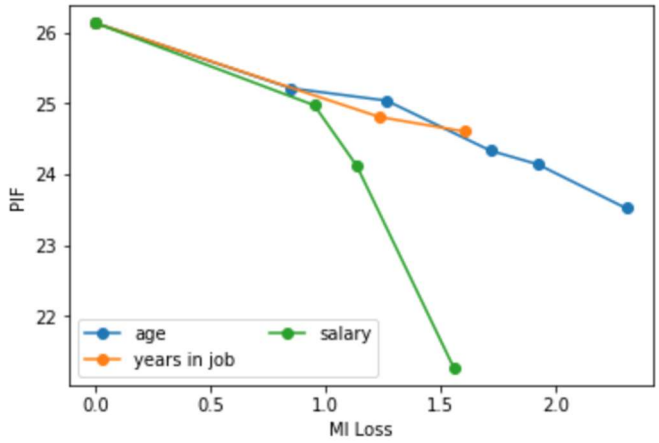
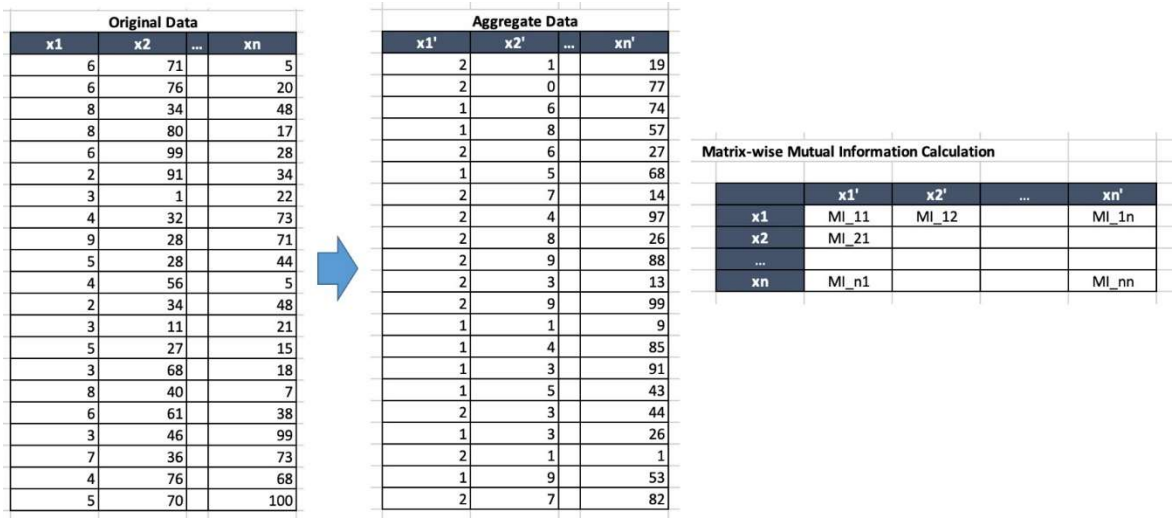


Figure 11. A measure of PIF versus MI of Aggregated Data Fields

If however dependence between features is known, then the potential exists to more carefully control the information loss as aggregation occurs. The concept of a *Matrix of Mutual Information* (MMI) was introduced which describes the matrix of mutual information loss between a feature in a dataset and an aggregated version of the same feature (see Figure 12). The MMI allows a more fine-grained analysis of which features to focus on for aggregation.



Original Data

x1	x2	...	xn
6	71		5
6	76		20
8	34		48
8	80		17
6	99		28
2	91		34
3	1		22
4	32		73
9	28		71
5	28		44
4	56		5
2	34		48
3	11		21
5	27		15
3	68		18
8	40		7
6	61		38
3	46		99
7	36		73
4	76		68
5	70		100

Aggregate Data

x1'	x2'	...	xn'
2	1		19
2	0		77
1	6		74
1	8		57
2	6		27
1	5		68
2	7		14
2	4		97
2	8		26
2	9		88
2	3		13
2	9		99
1	1		9
1	4		85
1	3		91
1	5		43
2	3		44
1	3		26
2	1		1
1	9		53
2	7		82

Matrix-wise Mutual Information Calculation

	x1'	x2'	...	xn'
x1	MI_11	MI_12		MI_1n
x2	MI_21			
...				
xn	MI_n1			MI_nn

Figure 12. Matrix of Mutual Information concept

The approach to using this information is to:

- Calculate the pairwise mutual information between features in the original dataset (as discussed in Section 3.2) to create a mutual information matrix (Original MI matrix)
- Aggregate each feature individually to produce a "Safer" dataset
- Calculate the pairwise mutual information between each feature in the original dataset and each feature in the aggregated dataset, the Matrix of Mutual Information (MMI)
- Calculate the total *MMI Loss* as the change in value for each feature pair between the Original MI matrix and the Matrix of Mutual Information.

Figure 13 outlines this process. Calculating the MMI Loss allows a means to track information loss as aggregation is applied to make datasets safer.

Original MI Matrix

	x1	x2	...	xn
x1	MI_11	MI_12		MI_1n
x2	MI_21			
...	..			
xn	MI_n1			MI_nn

Matrix of MI

	x1'	x2'	...	xn'
x1	MMI_11	MMI_12		MMI_1n
x2	MMI_21			
...	..			
xn	MMI_n1			MMI_nn

Matrix of MI Loss

	x1'	x2'	...	xn'
x1	MI_11-MMI_11	MI_12-MMI_12		MI_1n-MMI_1n
x2	MI_21-MMI_21			
...	..			
xn	MI_n1-MMI_n1			MI_nn-MMI_nn

Figure 13. Matrix of Mutual Information Loss

Returning to the example of dataset 8 (NSW Public Sector Workforce Synthetic dataset) as the features of "age", "salary" and "years in job" are independently aggregated, Figure 14 shows that the MMI loss when aggregating the salary feature is actually greater than that when aggregating other features. The implication is that, for a given level of PIF (RIG_{95}), aggregating salary leads to worse utility compared to aggregating other features in the dataset. This is in contrast to Figure 11 based on straight MI loss, which gave the misleading picture that aggregating salary would have the least impact on utility.

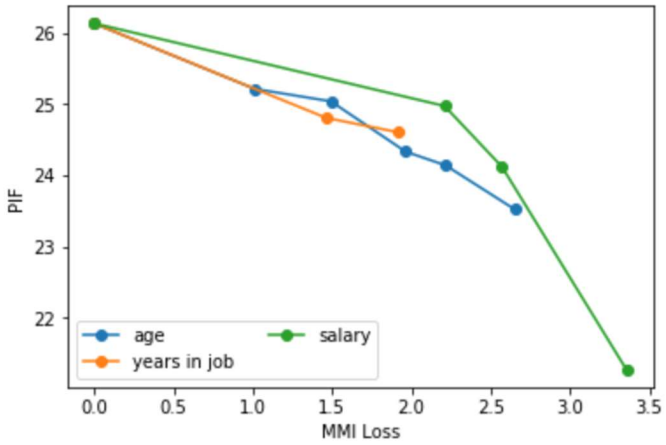


Figure 14. MMI loss for selected features in dataset 8 as each are independently aggregated

3.4 Implementation

Use of a standard PIF measure and standard thresholds allows the automated production of "safer" versions of a dataset when aggregation (or omission) are used as the means of reducing risk of reidentification. Figure 15 shows a simple feedback loop which does not consider any specific feature for preferential aggregation. The example method shown is "Least 2 values aggregated" which targets outlier values however many variations can be considered.

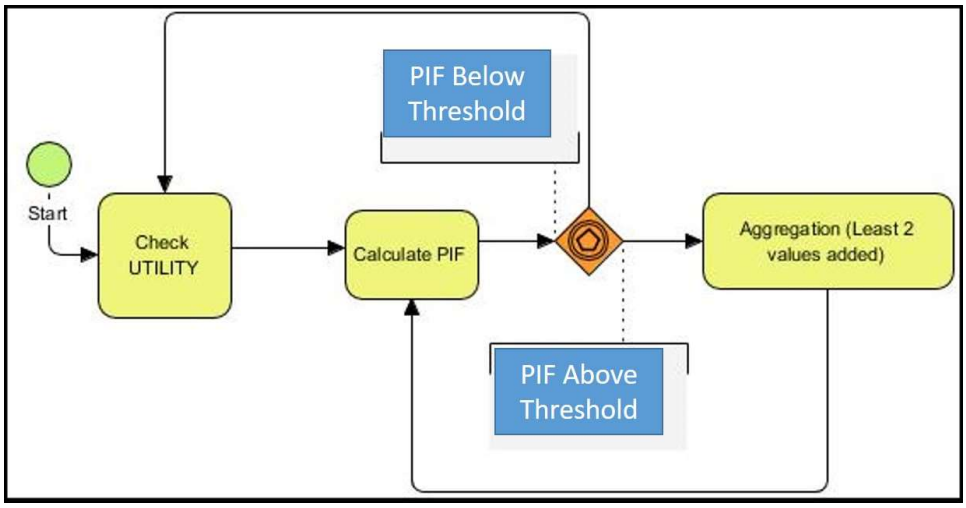


Figure 15. Automated PIF Evaluation

Based on the understanding of the loss of information, an example of a more sophisticated aggregation approach is shown in Figure 16. Many ways of aggregating (or omission) may be used to protect data, so this should be seen as an example only.

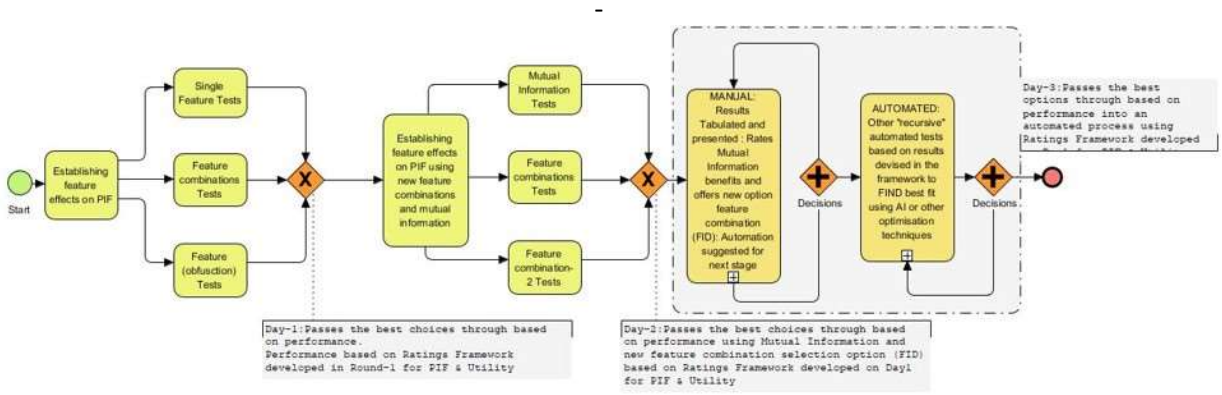


Figure 16. Example workflow for creation of "safe" datasets

The discussion above shows that knowledge of feature interdependence (and mutual information loss) has the potential to significantly improve the utility of datasets produced. Testing datasets at each of aggregation (or omission) of features may also improve dataset utility.

4 DEALING WITH TRAJECTORIES

One of the most significant challenges of working with people-centred data is dealing with longitudinal data or trajectories. When the history of appointments or admissions are linked to an individual, the ability to uniquely identify becomes very high.

4.1 Trajectory Flattening Techniques

The approach described in 0 was to "flatten" trajectories (see Figure 17) by exploring all possible sub-sequences for each possible feature (and all combinations of features) which can form a trajectory. The approach can very quickly become computationally intractable as many combinations of sub-sequence are identified. Also, the ability to identify unique trajectories readily becomes apparent based on simple parameters such as trajectory length or identification of a unique subsequence.

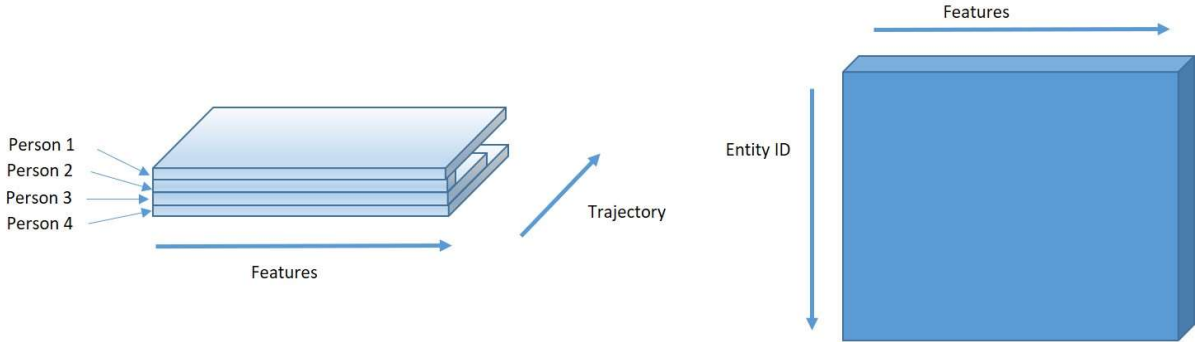


Figure 17. Trajectory decomposition

4.2 Depth Information Gain

The concept of Depth Information Gain (DIG) is analogous to cell information gain in that it considers values along the trajectory for each cell. It relies on the ability to identify a gain (loss) of information when the feature "trajectory" is examined. The challenge is to map a trajectory to a finite number of features to be examined as shown in Figure 18.

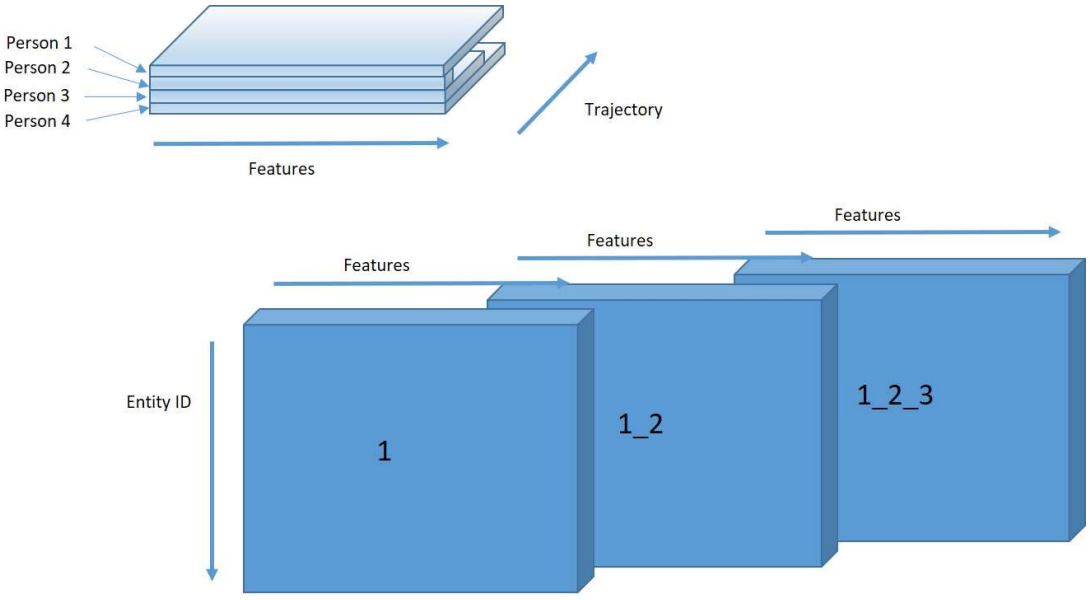


Figure 18. Decomposition of Trajectory into features

An evolution of the sub-sequence, the DIG approach considers the difference between steps in the sequence and identify the most unique transitions per stage as shown in Figure 18.

The process makes use of the CIG, which identified risk by cell, so that results from the trajectory analysis are comparable to the initial risk identification. The approach is computationally cheaper than vector embedding unique sequences and sub-sequences and should give a worst-case estimate.

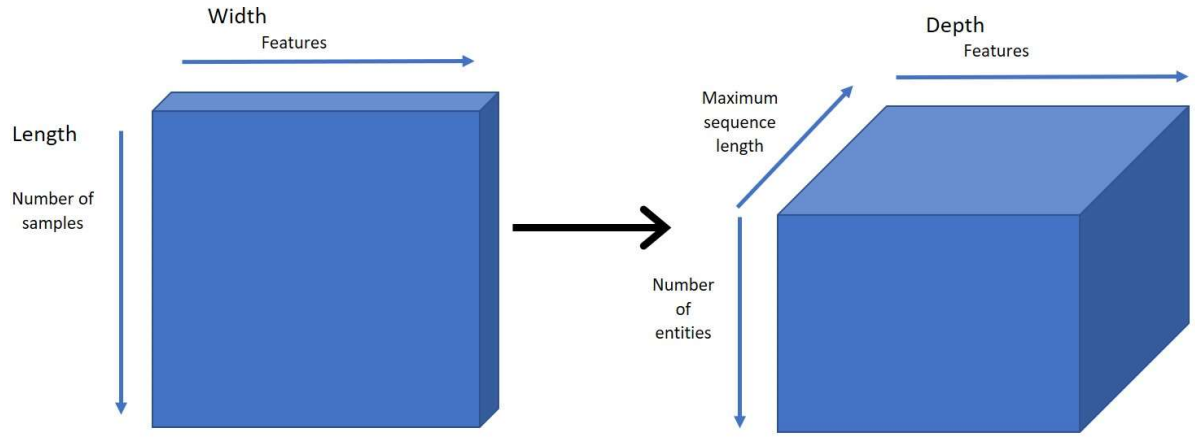


Figure 19. Trajectory decomposition

Using dataset 1 (Inmate Admissions) to describe the process (see Figure 19) :

1. Identify a feature which may contain trajectories of interest (such as prisoner I.D.)

2. Re-shape matrix from *sample x feature* to *I.D. x feature x sample*

   In the dataset considered, "length" is prisoner I.D, "width" remains as features, and depth becomes the samples themselves

3. For timesteps 2 through to the final step, concatenate preceding values for the equivalent cell (the preceding samples for each feature, for each I.D.)

   For example, if the first three samples for I.D. have values for feature "position" as 1, 2, and 3, the values become 1, 1_2, and 1_2_3. In this case, as CIG works by uniqueness, it does not matter that we change integers to strings – the uniqueness of the values in a particular timestep is what is being calculated.

4. Calculate CIG consecutively for the I.*D. x feature* matrix at the first timestep.

5. Using the original 2D (*sample x feature*) matrix, compare values for relevant rows, and store the maximum CIG value out of the previously stored value and the new calculation

6. Repeat steps 4-5 for the remaining timesteps

7. Repeat steps 1-6 for any other features which may form trajectories

Step 5 ensures that any given cell will report the highest risk for any sequence it is be part of (or its original risk, if that was equal to or higher than any sequence it is part of).

Figure 20 shows the evolution of values of the DIG at first step and the final step. Figure 21 shows the corresponding change in Mutual information at first step and after completion.

In these figures, the DIG baseline was performed on a subset of the dataset 1 (Inmate Admissions) with a max sequence length of 5 The DIG value after the final step was calculated after reducing all sequences to a max length of 2. The change in DIG shows that reducing the maximum sequence length reduced the number of unique sequences for *RACE* and *Inmate Status Code* – both of those showed a reduction in the maximum DIG. Some values increased by a small amount, due to the CIG calculation depending on the number of rows and features (in this case the number of rows would have been reduced). The change in MI showed that the distribution of data did not change much with the row removal but results for this would vary depending on the exact dataset used.

DIG Baseline

	Min	Q1	Avg	Med	Q3	Max
INMATEID	10.97	13.29	13.22	13.29	13.29	13.29
ADMITTED_DT	11.29	13.29	13.25	13.29	13.29	13.29
DISCHARGED_DT	0.91	0.91	7.06	5.66	13.29	13.29
RACE	0.91	0.91	1.43	1.18	1.18	13.29
GENDER	0.13	0.13	0.74	0.13	0.13	13.29
INMATE_STATUS_CODE	0.41	0.41	1.61	0.41	3.43	13.29
TOP_CHARGE	1.01	1.01	4.28	4.42	6.62	13.29

DIG After Final Step

	Min	Q1	Avg	Med	Q3	Max
INMATEID	11.36	13.28	13.22	13.28	13.28	13.28
ADMITTED_DT	11.28	13.28	13.25	13.28	13.28	13.28
DISCHARGED_DT	0.91	0.91	7.02	5.65	13.28	13.28
RACE	0.91	0.91	1.38	1.17	1.17	9.47
GENDER	0.13	0.13	0.69	0.13	0.13	13.28
INMATE_STATUS_CODE	0.41	0.41	1.56	0.41	3.46	12.28
TOP_CHARGE	1.00	1.00	4.25	4.47	6.46	13.28

Figure 20. DIG Baseline (Step 1) and After Processing

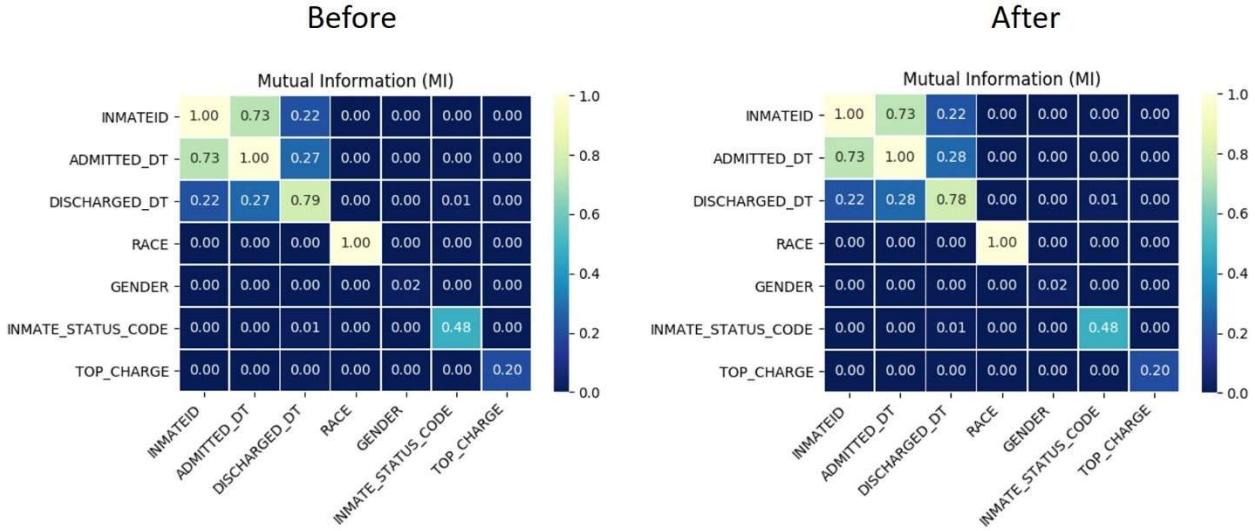


Figure 21. Mutual Information at Step 1 and After Processing

This approach reduces the complexity compares to full subsequence evaluation. An automated example implementation is shown in Figure 22.

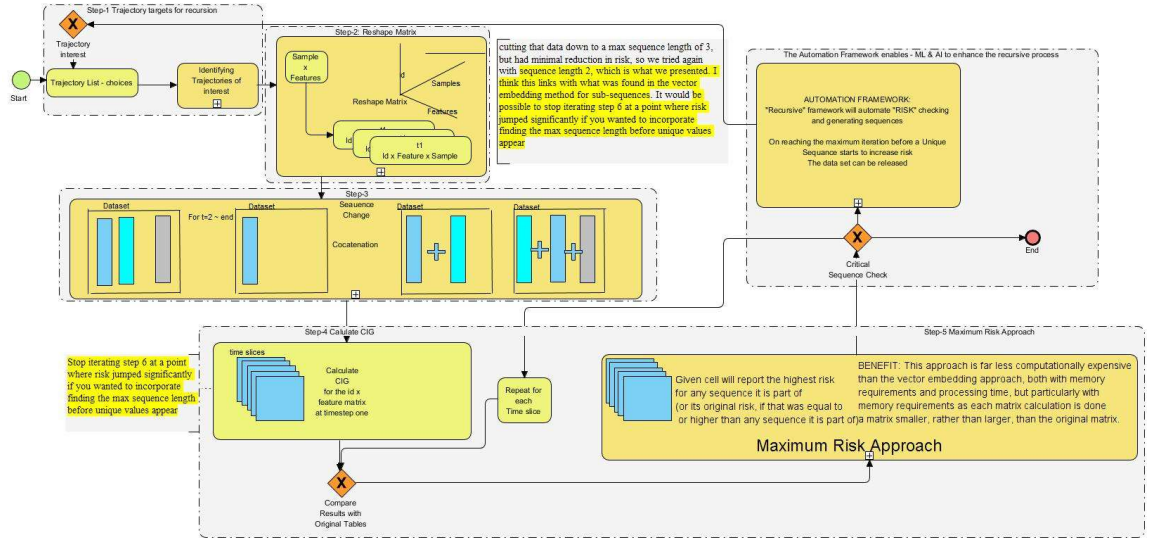


Figure 22. Example approach to dealing with trajectories

5. PROTECTING DATA THROUGH PERTURBATION

To this point, creating "safer" versions of a dataset has assumed aggregation or omission as the means of reducing the PIF. Adding random "noise" to a feature is a technique used by many agencies to make datasets safe(r) for public release.

5.1 Perturbation through Random Noise is Different

Adding random values to a data set (noise) with a strictly controlled distribution is a common technique for protecting data from the risk of reidentification. Adding noise with a Laplace distribution (see Figure 23) is a common approach as the random values can be tightly bound around a median value with the distribution used to change the level of protection

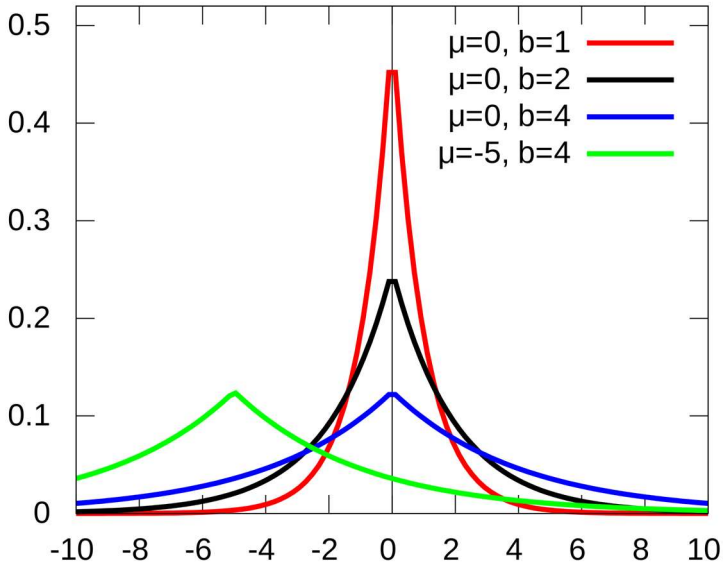


Figure 23. Laplace distributions based on central (μ) and deviation (b)

An immediate challenge posed by this approach is that every row (person) can readily become unique due to the random values applied to each feature. This renders the model of PIF, based on the smallest identifiable cohort, unable to address the uniqueness applies to random variations in feature values.

PIF and other entropy-based measures may also have certain weaknesses as privacy metrics, including strong outlier influence; reflect average rather than worst-case; and yield similar entropy-values for varied distributions, making it difficult to use as a 'metric' to compare different systems [6].

5.2 A Differential Privacy Approach

In recent years, differential privacy has been an active area of research. Differential privacy is a constraint which limits the disclosure of private information of records whose information is in the database. In simple terms, an algorithm is differentially private if an observer is unable to recognise the difference in its output of two datasets differing by an individual record, and is represented by the expression:

[6] Wagner, I., and Eckhoff, D. (2018). Technical privacy metrics: a systematic survey. ACM Computing Surveys (CSUR), 51(3), 57

$$\Pr[\mathcal{A}(D_1) \in S] \leq e^{\epsilon} \times \Pr[\mathcal{A}(D_2) \in S]$$

The randomised algorithm $\mathcal{A}$ is defined by a ratio of e^{ε} which represents a higher risk to privacy since there is a higher threshold of revealing differences between data sets D1 and D2.

Being founded on the notion of the difference made by the contribution of a single person or entity, the definition of a DP algorithm directly captures a very natural and intuitive notion of a mechanism for the release of a confidential dataset that preserves (within some specified tolerance, controlled by the parameter ε) the privacy of individual contributors to the dataset.

The Laplace Mechanism is the most well-known DP algorithm. It involves distorting the information contained in the input dataset by means of the injection of noise that is distributed according to the Laplace distribution. DP algorithms may be distinguished according to whether or not the amount of noise they inject depends on the input data set. The Laplace Mechanism is a data-independent algorithm.

The focus of this investigation is to use the notion of a DP algorithm to derive a metric that measures the relative safety of two given datasets. Here, a dataset is said to be safer than another dataset if the information it contains is more amenable to being released in a privacy-preserving manner than the information contained in the other dataset.

The hypothesis in question may be stated as follows: the less distortion that needs to be introduced into an input dataset by a data-dependent ε-differentially private algorithm (for some fixed value of ε), the safer the dataset.

In this case, a data-dependent DP algorithm is required for adding noise so that the noise reflects the properties of the data set. A potential candidate is the MWEM (Exponential Mechanism with the Multiplicative Weights) algorithm (Figure 24)[7]. MWEM operates on histogram representations of datasets. Starting from a uniform distribution and applying the Laplace Mechanism and another well-known DP algorithm called the Exponential Mechanism, it arrives at an approximate version of the input histogram, samples from which can be released. The released dataset is a distorted version of the input dataset, where the distortion is a consequence of injection of noise distributed according to the Laplace distribution.

Inputs: Data set B over a universe D, set Q of linear queries; Number of iterations $T \in \mathbb{N}$; Privacy parameter $\varepsilon > 0$.

Let n denote $\|B\|$, the number of records in B. Let A_0 denote n times the uniform distribution over D. For iteration $i = 1, ..., T$:

1. *Exponential Mechanism:* Sample a query $q_i \in Q$ using the Exponential Mechanism parametrized with epsilon value $\varepsilon/2T$ and the score function
$$s_i(B, q) = |q(A_{i-1}) - q(B)| .$$
2. *Laplace Mechanism:* Let measurement $m_i = q_i(B) + Lap(2T/\varepsilon)$.
3. *Multiplicative Weights:* Let A_i be n times the distribution whose entries satisfy
$$A_i(x) \propto A_{i-1}(x) \times \exp(q_i(x) \times (m_i - q_i(A_{i-1}))/2n) .$$

Output: $A = \text{avg}_{i<T} A_i$.

Figure 24. Description of the MWEM algorithm

[7] See M. Hardt, K. Ligett, F. McSherry, "A Simple and Practical Algorithm for Differentially Private Data Release", March 2012. Available online https://arxiv.org/pdf/1012.4763.pdf

Figure 25 shows a proposed methodology for determining the relative safety of two given datasets, D1 and D2:

1. Fix some value of ε.
2. Represent D1 and D2 as histograms (called H1 and H2, respectively).
3. Execute MWEM on H1, obtaining an output histogram H1'.
4. Calculate the KL-divergence Δ1 between H1 and H1'.
5. Execute MWEM on H2, obtaining an output histogram H2'.
6. Calculate the KL-divergence Δ2 between H2 and H2'.
7. Set the value of the metric for the safety of D1 relative to D2 to Δ2 / Δ1.

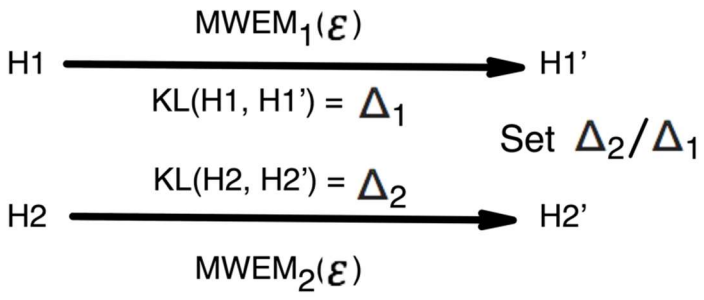


Figure 25. Differential Privacy Approach

A few points about the above approach which serves as a template:

- many executions of MWEM as feasible should be undertaken, and a mean and variance should be computed
- it could turn out that MWEM is not the most appropriate DP algorithm to employ, although it is a natural way to represent adding noise to histograms. It may be more appropriate to take an average of the metric values obtained for multiple DP algorithms.

In order to have some guidance on the selection of a suitable value of ε, one could fix a particular 'benchmark' value of the KL-divergence metric and take a ratio of the pair of ε values that are found to achieve that value.

It is important to note that the two datasets in question are arbitrary. In particular, the two datasets could be two variant 'privatisations' of a single confidential dataset (for example, a version obtained by aggregation and a version obtained by noise injection). Thus, one could use the methodology to determine which of the candidate privatisation strategies is safest for the given confidential dataset. In such a scenario, since the metadata for the two datasets are identical, the KL-divergence metric could be replaced by the mean squared error on some suitable fixed collection of histogram-level queries. The modified scenario is depicted in Figure 26.

The approach of generating datasets with different values of epsilon and comparing (see Figure 27) them allows a relative measure of privacy preservation to be explored as shown in Figure 28.

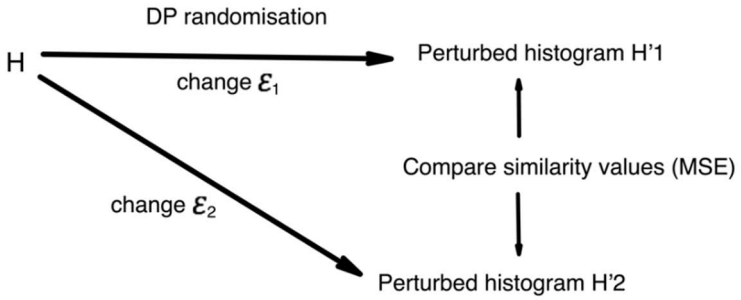


Figure 26 Comparing relative privacy between two perturbed datasets.

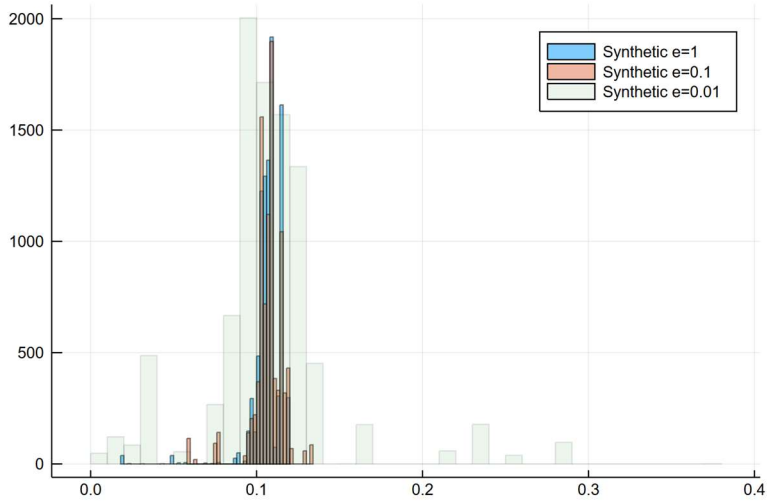


Figure 27. Synthetic datasets for different values of epsilon

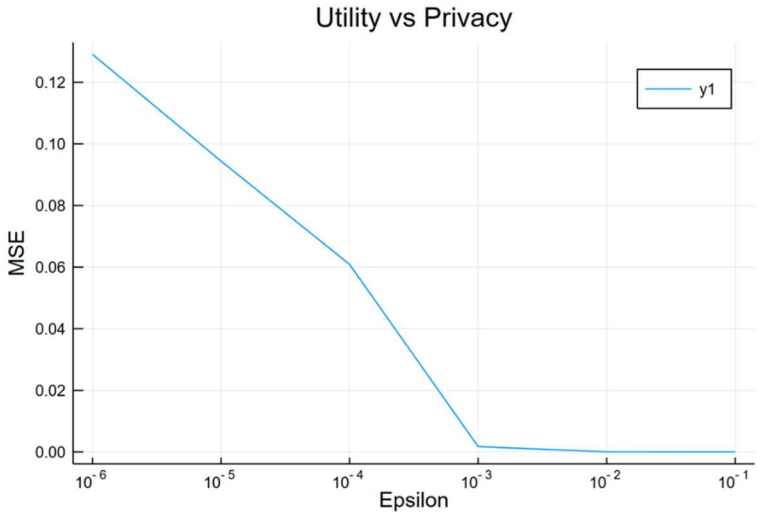


Figure 28. MSE as a measure of utility versus epsilon

The approach to privacy protection using differential privacy has the potential to be a complementary approach to the utility and PIF measures described above. Understanding the baseline PIF may allow a measure of differential privacy to be determined before perturbation. This is an area requiring further investigation.

6. CONCLUSIONS

We still have work to do.

This Directed Ideation demonstrated the feasibility of measures for Personal Information based on Information Theoretic approaches, which work with protection measures based on aggregation and omission. It also demonstrated the feasibility of measures of relative utility based on mutual information. During the course of the event, improvements were made to protection techniques based on identification of inter-dependence of features in a dataset.

The event also showed the potential of differential privacy-based approaches and the need for the personal information factor to evolve to deal with perturbation as a means of protection.

Dealing with Trajectories also proved to be a major challenge worthy of much further work.

During the write up of this report, a paper was published in Nature Communications [8] which provides a means to estimate the likelihood of a specific person to be correctly re-identified, even in a heavily incomplete dataset.

So, whilst incomplete, the work so far is useful even if in a limited scope of data sharing and with a specific "attacker" model in mind. Tools for utility, PIF (non-perturbed data), differential privacy (perturbed data), mutual information between features and mutual information loss all showed real promise for use in real-world systems.

Two of the major issues remaining are to operationalise the approaches using real datasets, and to link the measures back to the real-world challenge of privacy so that we can start to address the challenge of "reasonable likelihood" of reidentification.

[8] L. Rocher, J. M. Hendrickx & Y. de Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models", Nature Communications, July 2019. Available online https://www.nature.com/articles/s41467-019-10933-3

7. THANKS

Special thanks go to the contributors to the directed Ideation:

Georgina Kennedy, Brian Hope, Kelvin Ross, Arthur Street, Ollencio D'Souza, John Newman, Steve Woodyatt, James Kemp, Simone Reedy, Wanli Xue, Oisin Fitzgerald, Brian Thorne, Jim Basilakis, Leibo Liu, Tony Fish, Elliot Zhu, Gianpaolo Gioiosa and Geof Heydon.

Special thanks to Nick Rodwell, Michael Kam, Peter Chiu, Alex Byrganov, Jessica Kashro and Marc Portlock for their organising and coordinating expertise.

The 2018 ACS paper which this builds on was the culmination of more than two years work by a taskforce which included ACS, the NSW Data Analytics Centre (DAC), Standards Australia, the office of the NSW Privacy Commissioner, the NSW Information Commissioner, the Federal Government's Digital Transformation Agency (DTA), CSIRO, Data61, the Department of Prime Minister and Cabinet, the Australian Institute of Health and Welfare (AIHW), SA NT DataLink, South Australian Government, Victorian Government, West Australian Government, Queensland Government, the Communications Alliance, the Internet of Things Alliance Australia, DataSynergies, CreatorTech, Objective, EY, Microsoft, Clayton Utz and several other companies.

And finally, thanks to all others who have made, and continue to make, contributions and feedback.

APPENDIX A – February 2019 Event PIF Model

Cell Information Gain and Row Information Gain

One of the most important conclusions from the February 2019 was an information theoretic framework to quantify the information gain in the case of reidentification. The framework was grounded in concepts of information theory and cryptography.

The approach was motivated by the fact that with every dataset released, there is an increase in the information available about a person. However, not every reidentification event is of the same severity.

The framework then allows the user to:

- reason about risks on a per-feature (per-column in the table) basis,
- find risks of particular individuals (rows in the table),
- identify comparatively high-risk individuals,
- inform anonymisation efforts on where to focus, and
- compare different anonymisation strategies.

Quantifying Information

In designing risk metrics to be used on a number of different datasets, it is important that they be comparable. Otherwise, the values are not easily interpretable, and any attempts to set acceptable risk thresholds are doomed to fail. To ensure that the results are comparable, we use the same units to measure information gained by the attacker: bits.

Information theory is the field responsible for quantification of information. By building on it, we are leveraging well-known rigorous mathematical concepts. In information theory, the bit as a basic unit of information. For example, a coin toss is a binary choice where both options are equally likely, so it provides exactly one bit of information. If we have a biased coin, then the two outcomes are not equally likely and so the more likely outcome provides less than one bit of information. This makes intuitive sense since we already expected the more likely outcome: we do not learn as much if we are presented with information we already expect.

The approach was based on a K-L divergence calculation and produced a measure referred to as the Cell Information Gain (CIG), a Row Information Gain (RIG) and a Feature Information Gain (FIG).

Kullback–Leibler Divergence of Probability Distributions

A *probability distribution* is a list, possibly infinite, of possible choices for a value, along with the probability of each choice. For example, the probability distribution associated with a fair coin toss lists two outcomes: heads and tails. Each outcome has probability of one half.

The Kullback-Leibler divergence (KL-divergence) measures the information gain, in bits, when we update our belief from one probability distribution to another. If we are given a coin that may be biased, we might have a *prior* probability distribution that heads and tails are equally likely. This seems reasonable because we do not know how biased the coin is and in which direction. If we toss the coin 20 times and obtain heads 15 times, then our *posterior* probability distribution states that the coin's bias makes the probability of heads three quarters and the probability of tails one quarter. This updated belief represents 0.19 bits of information gain. This example is summarised in **Error! Reference source not found.**.

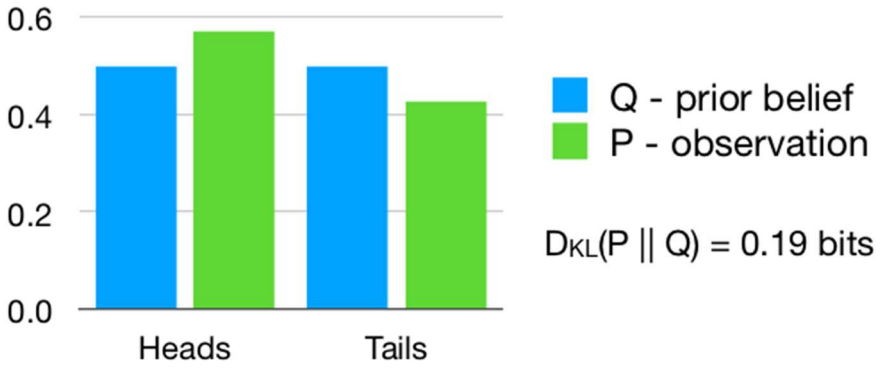


Figure 29. KL-Divergence of the coin toss example

Conversely, if upon investigation we find that our coin is unbiased, the KL-divergence of our prior and posterior probability distributions is 0 bits. This is because the probability, as estimated by us, of the outcome was unchanged.

When discussing reidentification, probability distributions are useful for modelling the information an attacker has about a person. The prior distribution represents the attacker's knowledge before they obtain the dataset. For example, if the attacker does not know a person's birthday, the prior would give each possible birthday equal probability. The posterior is what the attacker has been able to find by combining their existing information about a person with the information in the dataset. If the dataset permits our attacker to be sure about a person's birthday, then the posterior represents 8.5 bits of information gain. If they narrow it down to two equally likely options, then the information gain is 7.5 bits. If they learn nothing, then the KL-divergence of the prior and the posterior is zero.

The approach can therefore quantify information gain in the situation that the attacker does not become fully confident of a feature's value, but merely more confident.

Cell Information Gain

The team defined the *Cell Information Gain* (CIG) to quantify the reidentification risk for each piece of personal information. Every cell belongs to a row, and every row represents information about a person. We imagine that an attacker is attempting to reidentify a person to find the value of the cell whose CIG we are determining. We assume the attacker knows every feature of this person except this one cell. Its CIG is then defined as the KL-divergence of the attacker's prior and posterior beliefs for the true value of that cell.

The prior is the attacker's probability distribution for this cell before they attack the dataset. We often do not have access to this, so it is estimated (approximated) within the dataset by tallying the occurrences of every possible value of this feature across the entire dataset.

We calculate the posterior as well. Recall that at this point we have a particular person we imagine the attacker is targeting, and we have a vector of features for this person. To every person, or row, in the dataset we assign a probability that they are the person we seek to reidentify. For every possible value of our cell, we tally the probabilities of the people who have this value. This calculated posterior is compared with the prior to give us our CIG in bits.

Feature Information Gain

We can sum the Cell Information Gain for a feature to find the *Feature Information Gain* (FIG) for that feature. The FIG is a measure, in bits, of the reidentification risk of a feature. It can help us to identify the features that are the highest risk to include in a dataset. Of course, in any decision-making process the risk would be compared against the feature's utility when making the decision to include or exclude it.

Row Information Gain

We sum the Cell Information Gain for every row, or individual, of the table to obtain a *Row Information Gain* (RIG). It measures how susceptible a particular individual is to having their information revealed through reidentification in the dataset. The reason we are able to calculate a RIG by summing the CIGs is that we used consistent, comparable units for the CIG regardless of the feature.

Method

The overall approach was to:

- Remove columns (features) with unique identifiers such as license numbers, bank account numbers
- Estimate the distributions for each feature
- Calculate CIG values using K-L divergence
- Sum the CIG values per row to form the RIG (Row Information Gain) and per column to form FIG (Feature Information Gain) values
- Analyse RIG and FIG values to determine safety and inform next actions

The CIG is calculated (in bits) as

$$\begin{aligned} D_{\mathrm{KL}}(P \parallel Q) &= -\sum_{x \in \mathcal{X}} p(x) \log q(x) + \sum_{x \in \mathcal{X}} p(x) \log p(x) \\ &= \mathrm{H}(P, Q) - \mathrm{H}(P) \end{aligned}$$

Example: Consider a sample fictitious "Medical" dataset with CIG values shown in Figure 30.

In this dataset, Row 6 has large information gain for the "job" feature. All rows have large information gain for the "POSTCODE" feature.

Once identified, large CIG, RIG or FIG values can be altered or removed to reduce the PIF.

	gender	AGE	POSTCODE	blood_group	eye_color	job
0	0.736966	3.50706	7.67803	3.00353	2.33085	3.50535
1	1.32193	3.52638	8.39354	2.99185	2.31117	3.50535
2	1.32193	3.57562	8.83883	2.97789	2.31117	4.12917
3	1.32193	3.57562	7.16275	2.97789	2.32444	4.38644
4	1.32193	3.54684	11.2243	3.01561	2.33085	4.85394
5	1.32193	3.51905	5.38889	2.99185	2.33236	2.60658
6	0.736966	3.52638	11.2311	2.97789	2.33085	10.0613
7	1.32193	3.56803	5.47022	2.00597	2.33236	2.60658
8	0.736966	3.52638	7.87706	2.99185	2.31098	3.44733
9	1.32193	4.26248	8.72335	2.97996	2.33236	4.79756
10	1.32193	3.52768	8.27278	2.97996	2.33236	3.82016
11	1.32193	3.54684	6.83003	3.01168	2.33085	3.82016
12	1.32193	3.54684	8.25306	2.97789	2.31117	3.44733
13	0.736966	3.57562	11.4234	2.99185	2.33236	4.85394
14	1.32193	3.54684	7.57357	2.99185	2.33085	4.38644
15	0.736966	3.56803	4.87733	2.99185	2.31098	2.60658
16	1.32193	3.51518	7.68891	2.97996	2.31098	4.74733
17	1.32193	3.50706	0.22647	1.99582	2.33236	3.50535
18	0.736966	3.51905	5.21259	2.97789	2.31098	2.60658
19	0.736966	3.54684	11.3092	3.01561	2.31117	5.157

Figure 30. Example fictitious "Medical" dataset

Having such a fine-grain resolution of the information gain lets us reason over the dataset in different ways as described below.

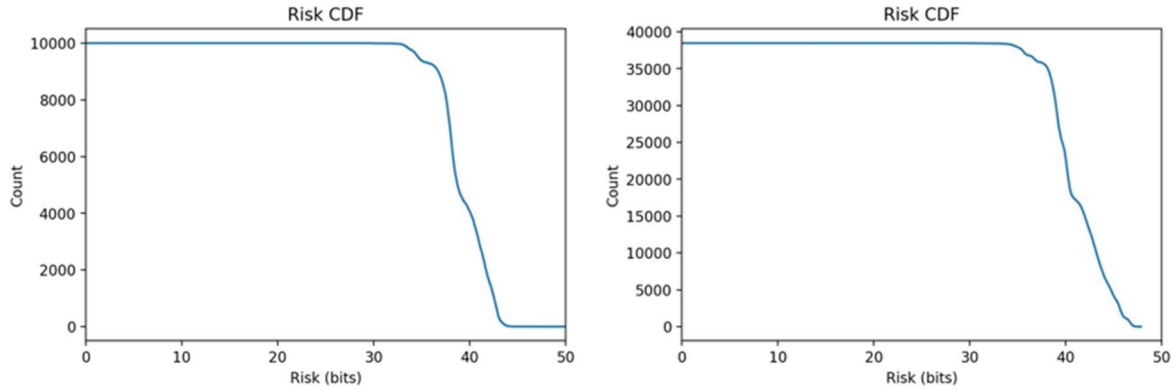


Figure 31. RIG score distribution by dataset. Against each risk (in bits) on the *x*-axis we plot the count of individuals whose own risk is higher than this.

Figure 31 shows the distributions of individuals' RIG scores in two datasets. The dataset on the left has RIG scores that are, on average, lower. However, that dataset has a very small number of individuals who are at an elevated risk of reidentification; this is visible in the long tail in the bottom right of the plot.

The team defined the quantity RIG_{95} to be the 95th percentile of the individuals' RIGs. This is one way of summarising the reidentification risk of the entire dataset in a single number. We also define RIG_{max} as the RIG of the most at-risk individual.

Flexibility of the Framework

The solid foundation of the team's framework makes it possible to extend and adapt to a wider range of use-cases.

Feature Accuracy

The framework for calculating the CIG allows inclusion of the level of noise in the information. This affects both of the posterior computation for the cell. The approach takes the uncertainty into account when assigning to each person the probability that they are the person being attacked. The approach also takes this into account when tallying those probabilities, combined with the feature values across the entire dataset, to produce a posterior for the cell. Generally, higher uncertainty in the data yields lower CIGs.

Incorporating Broader Knowledge About the Population

If more information is known about the distribution of a particular feature in the entire population rather than just the dataset, it is possible to base KL-divergence measure on these extended priors rather than on the dataset alone. This potentially allows for the data safety of low coverage datasets with unique values to be more appropriately measured.

Similarly, when creating "more safe" datasets from an example dataset, incorporating prior knowledge of how features are distributed across a population allows the approach to take into account broader knowledge about the data and reduce the impact of sampling on safety assessment.

	gender	birthdate	POSTCODE	blood_group	eye_color	icd_code
0	0.977816	4.09993	5.44488	0.417615	2.35482	7.80875
1	0.977816	4.09993	1.92583	0.417615	2.35482	7.80875
2	0.977816	4.09993	5.44488	0.417615	2.35482	7.80875
3	0.977816	4.09993	5.44488	0.417615	2.35482	7.80875
4	0.977816	4.09993	5.44488	0.417615	2.35482	7.80875
5	0.977816	2.1671	0.406784	0.417615	2.35482	2.85274
6	0.977816	4.09993	0.406784	0.417615	2.35482	2.85274
7	0.977816	1.34424	0.406784	0.417615	2.35482	2.85274
8	0.977816	1.69576	0.406784	0.417615	2.35482	2.85274
9	0.977816	2.14708	0.406784	0.417615	1.34966	2.85274
10	0.977816	1.69576	0.406784	0.417615	1.35751	2.85274
11	0.977816	2.20385	0.406784	0.417615	1.35742	2.85274
12	0.977816	2.17996	0.406784	0.417615	1.34605	2.85274
13	0.977816	2.1552	0.406784	0.417615	1.35751	2.85274
14	0.977816	2.5156	0.406784	0.417615	1.37974	2.85274
15	0.977816	2.99045	0.406784	0.417615	1.43826	2.85274
16	0.977816	1.68982	0.406784	0.417615	2.35482	2.85274
17	0.977816	2.15594	0.406784	0.417615	1.35751	2.85274
18	0.977816	1.79276	0.406784	0.417615	2.35482	2.85274
19	0.977816	2.20123	0.406784	0.417615	1.34966	2.85274

	gender	birthdate	POSTCODE	blood_group	eye_color	icd_code
0	0.977816	4.09993	5.44488	0.417615	2.35482	4.59894
1	0.977816	4.09993	1.92583	0.417615	2.35482	4.59894
2	0.977816	4.09993	5.44488	0.417615	2.35482	4.59894
3	0.977816	4.09993	5.44488	0.417615	2.35482	4.59894
4	0.977816	4.09993	5.44488	0.417615	2.35482	4.59894
5	0.977816	2.1671	0.406784	0.417615	2.35482	1.3596
6	0.977816	4.09993	0.406784	0.417615	2.35482	1.3596
7	0.977816	1.34424	0.406784	0.417615	2.35482	1.3596
8	0.977816	1.69576	0.406784	0.417615	2.35482	1.3596
9	0.977816	2.14708	0.406784	0.417615	1.34966	1.3596
10	0.977816	1.69576	0.406784	0.417615	1.35751	1.3596
11	0.977816	2.20385	0.406784	0.417615	1.35742	1.3596
12	0.977816	2.17996	0.406784	0.417615	1.34605	1.3596
13	0.977816	2.1552	0.406784	0.417615	1.35751	1.3596
14	0.977816	2.5156	0.406784	0.417615	1.37974	1.3596
15	0.977816	2.99045	0.406784	0.417615	1.43826	1.3596
16	0.977816	1.68982	0.406784	0.417615	2.35482	1.3596
17	0.977816	2.15594	0.406784	0.417615	1.35751	1.3596
18	0.977816	1.79276	0.406784	0.417615	2.35482	1.3596
19	0.977816	2.20123	0.406784	0.417615	1.34966	1.3596

Figure 32. Improved KL-divergence measures using knowledge of population distributions

Anonymisation Types

The technique for calculating CIG described here is agnostic to the kind of anonymisation used. A common technique for anonymisation is *k*-anonymity. It is obvious that the general scheme works in this case without modifications. Another approach may be to perturb the values before release. In this case we assign an accuracy to every feature and we take that into account as described above. The generality of this scheme comes from its solid grounding in probability theory and information theory.

Modelling Different Attackers

By default, the team modelled the attacker as very powerful assuming they known every feature of the person they are attempting to reidentify except from the one feature they are attempting to find. Nonetheless, different models for the attacker are also possible. These have connections to the Safe People aspect of the Five Safes framework.

In one model, the team assumed that the attacker knows *n* features of the person they are targeting. The feature they are attempting to find is not one of those *n*. Reasonably, an attacker that has less information about the person to begin with has less chance at reidentifying them. This is reflected by lower CIG (and consequently FIG and RIG) scores across the dataset.

Another possible model of the attacker assumes that they have some information but are not fully confident that it is correct. The level of confidence is a parameter that forms part of the assumptions in the approach.

Intuitively, if we assume that only Safe People are permitted to view the shared dataset, we may model the attacker as less powerful. This lets our safeguard be reflected in the reidentification risk calculation.

APPENDIX B – SAMPLE DATASETS

Dataset 1 – Inmate Admissions (United States open dataset)

Inmate admissions with attributes (race, gender, legal status, top charge). Record level with unique identifier of inmates. An inmate can have multiple charges, status, admission time, and discharged time.

301,748 rows and 7 columns,
148k unique inmate ID's.

Reference:

- Offence Charge Code: http://ypdcrime.com/penallawlist.php
- Full dataset and description: https://data.cityofnewyork.us/Public-Safety/Inmate-Admissions/6teu-xtgp

INMATEID	ADMITTED_DT	DISCHARGED_DT	RACE	GENDER	INMATE_STATUS_CODE	TOP_CHARGE
10001993	01/22/2018 06:32:26 PM		BLACK	M	DE	220.39
70983	1/02/2018 19:05	1/10/2018 20:17	UNKNOWN	M	DE	
2744	01/18/2018 05:40:04 PM		UNKNOWN	M	DE	140.2
20165517	1/09/2018 12:18		UNKNOWN	M	DE	110-120.05
20078557	01/15/2018 11:21:00 AM		BLACK	M	DE	155.25
20044863	1/07/2018 17:08		BLACK	M	DEP	120
111248	1/03/2018 16:17		BLACK	M	CS	215.5
20191524	01/25/2018 01:33:00 AM	01/29/2018 03:43:00 PM	BLACK	M	DE	
20190871	1/07/2018 12:20	1/08/2018 0:52	UNKNOWN	M	DE	
20129999	01/18/2018 11:09:36 AM		BLACK	M	DE	220.39
20150795	1/12/2018 19:40	01/18/2018 04:12:05 AM	UNKNOWN	M	DE	
20178129	01/31/2018 06:05:29 PM		UNKNOWN	F	DE	CO
43936	1/09/2018 3:33	1/09/2018 15:49	UNKNOWN	M	DE	
20191370	01/20/2018 08:03:00 PM		BLACK	M	DE	265.02
64122	1/05/2018 19:28	01/26/2018 09:29:01 AM	BLACK	M	CSP	120
165663	1/12/2018 11:50	1/12/2018 14:20	BLACK	M	DE	
4608	2/06/2016 2:13		BLACK	M	DEP	125.25
23108	1/06/2018 14:12	1/09/2018 22:45	BLACK	M	DE	
20190837	1/05/2018 20:08	1/06/2018 0:05	UNKNOWN	M	DE	

Figure 33. Sample of Inmate Admissions Dataset (United States open dataset)

Dataset 2 – Open Parking and Camera Violations (United States open dataset)

This dataset contains Open Parking and Camera Violations issued by the City of New York Record level on vehicle plate number with violation, and issue date. One vehicle plate can have multiple violations over time.

39.4m rows and 19 columns

Reference:

- Full dataset and description: https://data.cityofnewyork.us/City-Government/Open-Parking-and-Camera-Violations/nc67-uf89

Plate	State	License Type	Summons Number	Issue Date	Violation Time	Violation	Judgment Entry Date	Fine Amount	Penalty Amount	Interest Amount	Reduction Amount	Payment Amount	Amount Due	Precinct
GNY3760	NY	PAS	8653759098	4/05/2018	03:16P	SIDEWALK		115	0	0	0	115	0	110
8DC7395	MD	PAS	8653759104	4/05/2018	03:17P	SIDEWALK		115	0	0	0	115	0	110
GLR7577	NY	PAS	8602692663	5/04/2018	03:21P	REG. STICKER-EXPIRED/MISSING		65	0	0	0	65	0	122
HTT1406	NY	PAS	8661602403	05/14/2018	08:51A	NO PARKING-STREET CLEANING		45	0	0	0	45	0	94
2197026	IN	PAS	8602490288	03/13/2018	12:34P	NO STOPPING-DAY/TIME LIMITS		115	0	0	0	115	0	1
HSW8692	NY	PAS	8564044079	6/04/2018	07:31A	INSP STICKER-MUTILATED/C'FEIT	09/20/2018	65	60	0.5	0.22	125.28	0	112
LASTEVO	NY	SRF	8602692651	5/04/2018	03:19P	FAIL TO DSPLY MUNI METER RECPT		35	0	0	0	35	0	122
21974MG	NY	COM	8010541965	6/11/2015	01:27P	FAIL TO DISP. MUNI METER RECPT	10/01/2015	65	60	42.48	0	0	167.48	18
XCDE18	NJ	PAS	8600189070	5/04/2018	11:09A	NO STANDING-DAY/TIME LIMITS		115	30	0	0	145	0	18
86390MC	NY	COM	8600189032	5/04/2018	08:41A	FAIL TO DISP. MUNI METER RECPT		65	0	0	0	65	0	14
46052MG	NY	COM	8010542313	6/12/2015	12:18P	NO STANDING-DAY/TIME LIMITS	11/25/2015	115	60	55.7	0	0	230.7	14
89182MD	NY	COM	8010542854	06/16/2015	03:08P	NO STANDING-DAY/TIME LIMITS	10/01/2015	115	60	58.06	0	0	233.06	14
GXW2135	NY	PAS	8529199455	6/10/2018	11:44A	INSP. STICKER-EXPIRED/MISSING	09/27/2018	65	60	0.59	0.09	125.5	0	46
XGUG51	NJ	PAS	8688583882	05/21/2019	08:45A	NO PARKING-DAY/TIME LIMITS		65	10	0	10	65	0	20
GPS1075	NY	PAS	8602692808	5/04/2018	04:37P	FAIL TO DSPLY MUNI METER RECPT		35	0	0	35	0	0	122
HVJ7810	NY	PAS	8602692742	5/04/2018	04:10P	FAIL TO DSPLY MUNI METER RECPT		35	0	0	0	35	0	122
GWM1440	NY	PAS	8096539954	4/07/2017										
HRD6334	NY	PAS	8602692912	5/05/2018	08:47A	INSP. STICKER-EXPIRED/MISSING		65	0	0	0	65	0	121
T44HSG	NJ	PAS	8661602592	05/14/2018	11:49A	NO PARKING-STREET CLEANING		45	0	0	0	45	0	94
46323MG	NY	COM	8661602646	05/14/2018	12:04P	NO PARKING-STREET CLEANING		45	0	0	0	45	0	94
XDFT10	NJ	PAS	8602490665	03/14/2018	08:52A	NO STOPPING-DAY/TIME LIMITS		115	0	0	0	115	0	1
HKN9065	NY	PAS	8661602579	05/14/2018	11:46A	NO PARKING-STREET CLEANING		45	0	0	0	45	0	94
84969MJ	NY	COM	8529199674	6/11/2018	01:18P	DOUBLE PARKING		115	0	0	0	115	0	49
46052MG	NY	COM	8010543123	06/17/2015	12:02P	NO STANDING-DAY/TIME LIMITS	11/05/2015	115	60	56.56	0	0	231.56	14
HCX9470	NY	PAS	8529199753	6/11/2018	03:31P	INSP. STICKER-EXPIRED/MISSING		65	30	0	0	95	0	49

Figure 34. Sample Open Parking and Camera Violations (United States open dataset)

Dataset 3 – Air BNB Sydney Listings (commercial open dataset)

Publicly available information pooled by Inside Airbnb, with host ID, name, property listings, price, coordinates, text description, etc.

37,039 rows and 106 columns, with 27,335 unique host ids.

Reference:

- Data source: http://insideairbnb.com/get-the-data.html

id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_	availability_365
11156	An Oasis in	40855	Colleen		Sydney	-33.86917	151.22656	Private room	64	2	185	24/04/2019	1.61	1	352
12351	Sydney City	17061	Stuart		Sydney	-33.86515	151.1919	Private room	99	2	510	23/04/2019	4.76	2	200
14250	Manly Harb	55948	Heidi		Manly	-33.80093	151.26172	Entire home/apt	470	5	2	2/01/2019	0.05	2	40
15253	Stunning Pe	59850	Morag		Sydney	-33.88045	151.21654	Private room	110	2	321	20/04/2019	3.65	3	343
20865	3 BED HOU	64282	Fiona		Leichhardt	-33.85907	151.17275	Entire home/apt	450	7	16	3/01/2019	0.18	1	86
26174	COZY PRIVA	110561	Amanda		Woollahra	-33.88909	151.2594	Private room	61	1	45	29/03/2019	0.46	1	179
38073	Modern apa	103476	Prasanna		North Sydney	-33.83443	151.20887	Entire home/apt	159	2	63	16/09/2017	0.61	2	146
44545	Sunny Darli	112237	Atari		Sydney	-33.87996	151.21553	Entire home/apt	130	4	60	20/03/2019	0.58	1	0
57183	BONDI BEA	1623151	Susan		Waverley	-33.89185	151.27308	Entire home/apt	174	4	128	21/04/2019	1.26	1	140
58506	Studio Yind	279955	John		Mosman	-33.81927	151.23652	Entire home/apt	140	2	246	8/05/2019	2.41	1	246
58954	Christmas N	282630	Peter		Waverley	-33.89176	151.24259	Entire home/apt	1107	7	0			1	365
61721	2br Eclectic	299170	Eilish		Waverley	-33.8889	151.27726	Entire home/apt	244	4	25	26/02/2019	0.25	1	265
63795	Tree Tops R	311659	Tracey		Pittwater	-33.62612	151.33151	Entire home/apt	150	2	63	27/04/2019	0.63	1	306
65126	Large Garde	318390	Nicolette		Waverley	-33.88569	151.26886	Entire home/apt	150	5	11	21/04/2018	0.11	1	40
65635	Russell Hut	320878	Russell		Lane Cove	-33.81079	151.16072	Private room	54	1	165	19/04/2019	1.62	7	81
65857	Private Cou	322045	Jennifer		Sydney	-33.90396	151.19124	Private room	74	2	111	27/04/2019	2.81	1	7
66009	Comfort &	322887	Belinda		Woollahra	-33.88327	151.22725	Private room	100	3	1	28/02/2014	0.02	1	0
67112	Quiet base	160705	Liz		Marrickville	-33.915	151.1403	Private room	74	3	22	17/04/2015	0.22	1	363
68999	A little bit o	333581	Brian		Hornsby	-33.7299	151.05138	Private room	89	3	46	29/01/2019	0.48	1	91
69121	northern be	345292	Pamela		Warringah	-33.71249	151.29842	Entire home/apt	110	21	0			1	131

Figure 35. Sample of Air BNB Sydney Listings (commercial open dataset)

Dataset 4 - NYC Green Taxi Trip Data (United States open dataset)

The green taxi trip records include fields pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.

8.81m rows and 19 columns.

Reference:

- Full dataset and description: https://data.cityofnewyork.us/Transportation/2018-Green-Taxi-Trip-Data/w7fs-fd9i

VendorID	lpep_pickup_datetime	lpep_dropoff_datetime	store_and_fwd_flag	RatecodeID	PULocationID	DOLocationID	passenger_count	trip_distance	fare_amount	extra	mta_tax	tip_amount	tolls_amount	ehail_fee	improvement_surcharge	total_amount	payment_type	trip_type
2	1/01/2018 0:18	1/01/2018 0:24	N	1	236	236	5	0.7	6	0.5	0.5	0	0		0.3	7.3	2	1
2	1/01/2018 0:30	1/01/2018 0:46	N	1	43	42	5	3.5	14.5	0.5	0.5	0	0		0.3	15.8	2	1
2	1/01/2018 0:07	1/01/2018 0:19	N	1	74	152	1	2.14	10	0.5	0.5	0	0		0.3	11.3	2	1
2	1/01/2018 0:32	1/01/2018 0:33	N	1	255	255	1	0.03	-3	-0.5	-0.5	0	0		-0.3	-4.3	3	1
2	1/01/2018 0:32	1/01/2018 0:33	N	1	255	255	1	0.03	3	0.5	0.5	0	0		0.3	4.3	2	1
2	1/01/2018 0:38	1/01/2018 1:08	N	1	255	161	1	5.63	21	0.5	0.5	0	0		0.3	22.3	2	1
2	1/01/2018 0:18	1/01/2018 0:28	N	1	189	65	5	1.71	8.5	0.5	0.5	0	0		0.3	9.8	2	1
2	1/01/2018 0:38	1/01/2018 0:55	N	1	189	225	5	3.45	14.5	0.5	0.5	3.16	0		0.3	18.96	1	1
2	1/01/2018 0:05	1/01/2018 0:18	N	1	129	82	1	1.61	10	0.5	0.5	0	0		0.3	11.3	2	1
2	1/01/2018 0:35	1/01/2018 0:42	N	1	226	7	1	1.87	7.5	0.5	0.5	0	0		0.3	8.8	2	1
2	1/01/2018 0:21	1/01/2018 0:39	N	1	145	129	2	4.12	16.5	0.5	0.5	0	0		0.3	17.8	2	1
2	1/01/2018 0:56	1/01/2018 1:04	N	1	7	223	2	1.22	7	0.5	0.5	0	0		0.3	8.3	2	1
2	1/01/2018 0:11	1/01/2018 0:30	N	1	255	189	1	4.67	17	0.5	0.5	0	0		0.3	18.3	2	1
2	1/01/2018 0:57	1/01/2018 1:12	N	1	97	188	1	2.71	11.5	0.5	0.5	3.84	0		0.3	16.64	1	1
2	1/01/2018 0:36	1/01/2018 0:51	N	1	244	75	2	6.01	19	0.5	0.5	4	0		0.3	24.3	1	1
1	1/01/2018 0:07	1/01/2018 0:15	N	1	225	37	1	1.9	8	0.5	0.5	3	0		0.3	12.3	1	1
1	1/01/2018 0:25	1/01/2018 0:42	N	1	36	145	2	4.3	15.5	0.5	0.5	3.35	0		0.3	20.15	1	1
1	1/01/2018 0:42	1/01/2018 1:00	N	1	145	173	1	6.9	22	0.5	0.5	0	0		0.3	23.3	1	1
2	1/01/2018 0:06	1/01/2018 0:08	N	1	49	49	1	0.3	3.5	0.5	0.5	0	0		0.3	4.8	2	1
2	1/01/2018 0:34	1/01/2018 0:52	N	1	40	113	1	4.47	16.5	0.5	0.5	3.56	0		0.3	23.31	1	1
1	1/01/2018 0:25	1/01/2018 0:28	N	1	179	7	1	0.5	4.5	0.5	0.5	0	0		0.3	5.8	1	1
2	1/01/2018 0:36	1/01/2018 0:51	N	1	7	193	1	1.82	9	0.5	0.5	0	0		0.3	10.3	1	1
2	1/01/2018 0:53	1/01/2018 1:26	N	1	97	74	1	11.79	36	0.5	0.5	7.46	0		0.3	46.71	1	1
1	1/01/2018 0:11	1/01/2018 0:22	N	1	255	112	1	1.9	9	0.5	0.5	3.05	0		0.3	13.35	1	1
1	1/01/2018 0:40	1/01/2018 1:01	N	1	255	28	1	10.3	29	0.5	0.5	5	0		0.3	35.3	1	1
2	1/01/2018 0:15	1/01/2018 0:25	N	1	80	80	1	1.66	8.5	0.5	0.5	1.96	0		0.3	11.76	1	1
2	1/01/2018 0:35	1/01/2018 0:48	N	1	255	232	1	2.91	12	0.5	0.5	3.32	0		0.3	16.62	1	1
2	1/01/2018 0:55	1/01/2018 1:28	N	1	256	50	1	6.09	25	0.5	0.5	6.58	0		0.3	32.88	1	1
2	1/01/2018 0:41	1/01/2018 0:56	N	1	179	75	5	5.3	17	0.5	0.5	3.66	0		0.3	21.96	1	1
2	1/01/2018 0:36	1/01/2018 0:44	N	1	41	75	1	1.63	8	0.5	0.5	1.86	0		0.3	11.16	1	1
2	1/01/2018 0:48	1/01/2018 0:51	N	1	75	74	1	0.91	4.5	0.5	0.5	0	0		0.3	5.8	2	1
2	1/01/2018 0:56	1/01/2018 1:00	N	1	74	74	2	0.92	5	0.5	0.5	0	0		0.3	6.3	1	1
2	1/01/2018 0:27	1/01/2018 0:34	N	1	7	223	1	0.98	6	0.5	0.5	0	0		0.3	7.3	2	1
2	1/01/2018 0:41	1/01/2018 0:52	N	1	179	7	1	1.42	9	0.5	0.5	0	0		0.3	10.3	2	1

Figure 36. Sample of NYC Green Taxi Trip Data (United States open dataset)

Dataset 5 - ATO Taxation Individual Statistics (Australian open datasets)

Aggregated individual taxation statistics by industry consisting financial year 2013-14, 2014-15, 2015-16, and 2016-17 (four separate datasets combined). Included are description of industry, amount of tax, taxable income, medicare levy and superannuation.

2,204 rows and 138 columns

Reference:

- FY 13-14: https://data.gov.au/dataset/ds-dga-25e81c18-2083-4abe-81b6-0f530053c63f
- FY 2014-15: https://data.gov.au/dataset/ds-dga-5c99cfed-254d-40a6-af1c-47412b7de6fe
- FY 2015-16: https://data.gov.au/dataset/ds-dga-d170213c-4391-4d10-ac24-b0c11768da3f
- FY 2016-17: https://data.gov.au/dataset/ds-dga-540e3eac-f2df-48d1-9bc0-fbe8dfec641f

Financial Year	Broad Industry[1,4,5]	Fine Industry[1]	Number of individuals	Taxable income or loss[3] no.	Taxable income or loss[3] $	Gross tax no.	Gross tax $	Medicare levy no.	Medicare levy $	Medicare levy surcharge no.	Medicare levy surcharge $
2013-14	A.Agriculture, Forestry and Fishing	01110 Nursery Production (Under Cover)	506	499	22,316,956	350	4,599,628	275	286,011	6	5,650
2013-14	A.Agriculture, Forestry and Fishing	01120 Nursery Production (Outdoors)	643	627	45,893,760	433	14,072,454	354	643,151	9	7,939
2013-14	A.Agriculture, Forestry and Fishing	01130 Turf Growing	156	153	6,662,180	118	1,263,965	95	87,656	2	2,752
2013-14	A.Agriculture, Forestry and Fishing	01140 Floriculture Production (Under Cover)	85	85	3,077,301	57	551,908	43	39,228	0	0
2013-14	A.Agriculture, Forestry and Fishing	01150 Floriculture Production (Outdoors)	245	238	10,775,869	156	2,863,319	112	148,547	0	0
2013-14	A.Agriculture, Forestry and Fishing	01210 Mushroom Growing	71	68	3,737,521	48	919,301	37	50,458	1	1,356
2013-14	A.Agriculture, Forestry and Fishing	01220 Vegetable Growing (Under Cover)	527	513	15,944,299	389	2,258,220	259	179,171	6	6,362
2013-14	A.Agriculture, Forestry and Fishing	01230 Vegetable Growing (Outdoors)	1,262	1,228	46,963,532	840	10,342,358	607	625,745	12	14,947
2013-14	A.Agriculture, Forestry and Fishing	01310 Grape Growing	2,088	2,012	214,371,317	1,554	71,694,341	1,334	3,189,426	18	15,876
2013-14	A.Agriculture, Forestry and Fishing	01320 Kiwifruit Growing	14	14	448,166	11	51,717	8	5,241	0	0
2013-14	A.Agriculture, Forestry and Fishing	01330 Berry Fruit Growing	110	106	4,816,615	71	1,129,197	58	67,411	4	16,440
2013-14	A.Agriculture, Forestry and Fishing	01340 Apple and Pear Growing	96	88	4,973,783	65	1,261,023	50	67,827	0	0
2013-14	A.Agriculture, Forestry and Fishing	01350 Stone Fruit Growing	174	165	10,469,548	121	2,980,432	95	149,485	3	3,770
2013-14	A.Agriculture, Forestry and Fishing	01360 Citrus Fruit Growing	271	258	14,242,690	185	3,728,969	148	199,953	1	961
2013-14	A.Agriculture, Forestry and Fishing	01370 Olive Growing	458	449	58,189,035	382	18,763,825	358	858,517	8	13,461
2013-14	A.Agriculture, Forestry and Fishing	01390 Other Fruit and Tree Nut Growing	1,963	1,904	224,427,388	1,609	72,394,523	1,435	3,311,293	28	39,799
2013-14	A.Agriculture, Forestry and Fishing	01410 Sheep Farming (Specialised)	3,463	3,324	152,638,265	2,291	38,714,288	1,819	2,195,921	35	60,308
2013-14	A.Agriculture, Forestry and Fishing	01420 Beef Cattle Farming (Specialised)	19,349	18,436	979,451,517	12,296	287,801,956	10,058	14,968,755	226	355,374
2013-14	A.Agriculture, Forestry and Fishing	01430 Beef Cattle Feedlots (Specialised)	70	69	4,072,901	49	1,426,946	42	71,597	0	0
2013-14	A.Agriculture, Forestry and Fishing	01440 Sheep-Beef Cattle Farming	6,122	5,778	579,824,947	3,859	218,603,949	3,119	8,871,375	75	131,855
2013-14	A.Agriculture, Forestry and Fishing	01450 Grain-Sheep or Grain-Beef Cattle Farming	3,806	3,541	153,597,008	2,512	43,217,865	2,050	2,503,500	67	98,711
2013-14	A.Agriculture, Forestry and Fishing	01460 Rice Growing	113	104	4,441,186	86	1,127,709	69	71,980	2	3,029

Figure 37. Sample of ATO Taxation Individual Statistics (Australian open datasets)

Dataset 6 - Synthetic NAPLAN Test Result Data (Synthetic dataset)

Randomly generated unit record level of student performance on the NAPLAN test. Each record has a student's name, country of birth, year level, one parent's occupation group, School ID, and the test results in the form of bands. The randomly generated test result consists of reading, spelling, grammar and punctuation, writing, and numerical literacy. Data is randomly generated however adheres to the major statistical properties of the original dataset.

Reference:

- More about NAPLAN test: https://www.nap.edu.au/naplan

SchoolID	Surname	First_Name	Gender	DOB	Year_Level	Student_Country_of_birth	Parent1_Occup_Group	readband	splband	grpnband	writband	numband
283	Montoya	Kim	2	25/12/2008	5	1101	2	4	5	5	6	5
2701	Myers	Jason	1	23/01/2009	5	1101	4	7	7	7	6	6
770	Grant	Sharon	2	7/01/2007	7	1100	1	5	6	6	5	5
443	Rush	John	1	16/12/2010	3	1101	4	1	4	1	4	1
504	Gonzalez	Robert	1	1/08/2006	7	1101	4	6	5	6	6	8
2417	Cole	Sabrina	2	1/09/2010	3	1101	4	4	6	6	6	5
872	Scott	James	1	3/04/2011	3	1101	2	4	5	3	5	5
1405	Scott	Cheryl	2	7/05/2009	5	1101	9	6	6	6	5	5
1150	Perez	Michael	1	31/05/2007	7	1101	3	5	6	7	7	8
537	Webb	Sharon	2	30/11/2004	9	1101	3	8	8	8	7	8
1739	Foster	David	1	15/07/2004	9	1101	2	9	9	7	8	8
420	Peterson	Terry	1	12/09/2006	7	1101	3	6	5	5	4	6
2483	Gray	Jody	2	20/05/2009	5	1101	1	6	8	7	6	6
2468	Patel	Franklin	1	9/04/2011	3	2100	9	3	5	5	5	4
1284	Ibarra	Justin	1	18/07/2004	9	1101	3	9	9	10	8	10
1661	Cole	Jessica	2	5/06/2007	7	1101	2	6	6	6	5	5
1225	Gould	Nicole	2	8/03/2005	9	1101	8	6	7	6	7	7
192	Orozco	Christina	2	2/12/2010	3	1101	9	6	6	6	6	6
2378	Morris	Albert	1	24/02/2007	7	1101	3	5	5	5	5	5

Figure 38. Sample of Synthetic NAPLAN Test Result Data (Synthetic dataset)

Dataset 7 - Synthetic Hospital Admissions Data (Synthetic dataset)

Randomly generated dataset with fields including personal information (name, address, DOB, occupation) as well as medical diagnosis from ICD10 (International Classification of Diseases 10th Revision)[9]. Record level of individuals admitted to the hospital with diagnosis details, date of birth, gender, occupation, and address. Each individual synthetic patient has a trajectory of different visit time and diagnosis.

1.4m Rows with 14 columns
96,724 unique synthetic patient ID's

Reference:

- Prevalence of medical condition in Australia is generated from: http://ghdx.healthdata.org/gbd-results-tool

name	gender	patientid	birthdate	countryofbirth	address	blood_group	eye_color	job	company	visittime	age	diagnosis_code	diagnosis_desc
Catherine Phillips	F	287-86-8304	9/09/1928	Jordan	14 Cline GateH	AB+	Blue	Microbiolc	Donaldson	1/01/1952 1:35	23	515	Asthma
Thomas Jones	M	533-49-6215	25/06/1933	Holy See (Vatican City State)	1 / 51 Michael I	B+	Hazel	Student	NA	1/01/1952 5:05	18	668	Other skin and subcutaneous diseases
Courtney Huber	F	692-26-4478	25/04/1951	Philippines	27 / 2 Jillian Cr	B-	Brown	NA	NA	1/01/1952 6:30	0	328	Upper respiratory infections
Thomas Petty	M	419-64-4893	12/09/1948	Saint Lucia	34 / 93 Taylor (	AB-	Hazel	NA	NA	1/01/1952 13:55	3	681	Caries of deciduous teeth
Mathew Phillips	M	831-18-8881	28/07/1930	Thailand	Level 0 253 Sm	O+	Green	Student	NA	1/01/1952 14:14	21	562	Opioid use disorders
Savannah Hicks	F	801-23-5015	19/07/1933	Solomon Islands	Flat 75 3 Hart R	O+	Grey	Student	NA	1/01/1952 14:29	18	630	Low back pain
William Patton	M	225-34-6686	31/10/1934	Peru	Level 0 8 Weiss	A+	Hazel	Student	NA	1/01/1952 17:22	17	548	Tension-type headache
Karen Davis	F	382-40-5508	4/08/1942	Bulgaria	Flat 32 580 Eliz	B-	Brown	Student	NA	1/01/1952 18:01	9	681	Caries of deciduous teeth
James Lyons	M	665-99-2774	29/11/1935	Estonia	8 Campbell Bra	O-	Green	Student	NA	1/01/1952 18:18	16	707	Other exposure to mechanical forces
Tasha Davis	F	263-44-9533	14/05/1944	Somalia	080 Matthew R	A+	Blue	Student	NA	1/01/1952 18:28	7	682	Caries of permanent teeth
Tristan Fisher	M	543-13-6020	28/03/1927	Nepal	Suite 640 9 Buc	AB-	Blue	Illustrator	Hampton	1/01/1952 18:40	24	547	Migraine
William Garcia	M	664-65-6728	15/12/1941	French Southern Territories	373 Wilson Rar	O+	Brown	Student	NA	1/01/1952 19:09	10	682	Caries of permanent teeth
Derek Glass	M	637-38-4799	9/03/1939	Macao	Apt. 303 7 Wilk	B+	Hazel	Student	NA	1/01/1952 21:04	12	668	Other skin and subcutaneous diseases
Nancy Harvey	F	702-15-6168	6/06/1934	Vanuatu	5 Guerra Mews	B-	Blue	Student	NA	1/01/1952 22:32	17	694	Other road injuries
Ricardo Perez	M	331-10-9361	28/10/1923	Liberia	616 Jackson Hil	O-	Blue	Estate age	Davis, Wils	1/01/1952 23:06	28	659	Fungal skin diseases
Jonathan Silva	M	580-87-1961	3/08/1923	Antigua and Barbuda	9 Mendoza Ave	A-	Grey	Arboricult	Reyes-Mo	1/01/1952 23:32	28	682	Caries of permanent teeth
James Rivera	M	418-17-8845	1/07/1942	Algeria	Unit 36 316 De	AB-	Brown	Student	NA	1/01/1953 0:22	10	389	Vitamin A deficiency
Cindy Chang	F	068-11-4230	5/12/1936	Italy	1 / 45 Daniel Lo	O-	Hazel	Student	NA	1/01/1953 1:25	16	685	Other oral disorders
Michael Brown	M	464-95-8653	25/06/1946	Finland	6 Howe Terrace	AB+	Brown	Student	NA	1/01/1953 8:25	6	681	Caries of deciduous teeth
Ashley Reyes	F	647-43-3234	6/08/1924	Honduras	7 Marks NookP	AB+	Green	Scientist, p	Le, Brown	1/01/1953 10:48	28	548	Tension-type headache
Robert Fuller	M	112-66-3822	16/05/1949	Benin	94 Jill CorsoAsh	B+	Brown	NA	NA	1/01/1953 12:22	3	838	Sickle cell trait
Caitlin Ramirez	F	760-48-3377	7/10/1938	Sudan	40 / 674 Alvara	AB+	Grey	Student	NA	1/01/1953 13:18	14	668	Other skin and subcutaneous diseases
Carlos Foster	M	847-54-1496	1/12/1934	Vanuatu	Apt. 259 268 M	A-	Hazel	Student	NA	1/01/1953 14:05	18	571	Anxiety disorders
Melanie York	F	015-18-7147	1/02/1924	Panama	837 Leonard Ov	A-	Brown	Museum/g	Faulkner P	1/01/1953 14:42	28	609	Premenstrual syndrome

Figure 39. Sample of Synthetic Hospital Admissions Data (Synthetic dataset)

[9] See https://www.cdc.gov/nchs/icd/icd10cm.htm (accessed July 2019).

Dataset 8 – Synthetic NSW People Matter Employee Survey (PMES) (Synthetic Dataset)

Randomly generated dataset with fields including demographic attributes of the survey respondents (education level, age group, disability status, employment status, gender, LGBTI status, and ethnical diversity) along with the Likert scale responses to the survey questions.

180,000 rows with 117 columns

Reference:

- More information about PMES: https://www.psc.nsw.gov.au/reports---data/people-matter-employee-survey

ID	ATSI_Status	Age_Group	Current_Role_Years_Employed	Disability_Status	Education	Employment_Status	Gender	Gross_Salary	LGBTI_Status	LOTE_Status
1		40 - 44	1 - 2 years	No	Bachelor Degree level	Ongoing/Permanent (other than senior executive)	Male	$121,917 - $140,995	No	
2	No	35 - 39	2 - 5 years	No	Graduate Diploma or Graduate Certificate level	Ongoing/Permanent (other than senior executive)	Female	$183,300 - $261,450	No	No
3	No	45 - 49	Less than 1 year	No	Bachelor Degree level		Male	$121,917 - $140,995	No	No
4	No	50 - 54	10 - 20 years	No	Graduate Diploma or Graduate Certificate level	Ongoing/Permanent (other than senior executive)	Female	$183,300 - $261,450	No	No
5	No	65+		No	Prefer not to say		Female	$183,300 - $261,450	No	No
6	Prefer not to say	45 - 49	More than 20 years	No	Advanced Diploma or Diploma level	Ongoing/Permanent (other than senior executive)	Female	$183,300 - $261,450	No	Prefer not to say
7	No	40 - 44	2 - 5 years	No	Less than year 12 or equivalent	Ongoing/Permanent (other than senior executive)	Female	$183,300 - $261,450	No	Yes
8	No	50 - 54	10 - 20 years	No	Bachelor Degree level	Ongoing/Permanent (other than senior executive)	Female	$183,300 - $261,450	No	No
9	No	45 - 49	Less than 1 year	No	Prefer not to say	Ongoing/Permanent (other than senior executive)	Female	$183,300 - $261,450	No	No
10	No	45 - 49	1 - 2 years	No	Certificate level, including trade	Ongoing/Permanent (other than senior executive)	Female	$183,300 - $261,450	No	Yes
11	Yes	60 - 64	2 - 5 years	No	Bachelor Degree level	Labour hire	Female	$183,300 - $261,450	No	No
12	No	30 - 34	More than 20 years	No	Bachelor Degree level	Ongoing/Permanent (other than senior executive)	Female	$183,300 - $261,450	No	Prefer not to say
13	No	55 - 59	2 - 5 years	No	Less than year 12 or equivalent	Ongoing/Permanent (other than senior executive)	Male	$121,917 - $140,995	No	No
14	No	30 - 34	10 - 20 years	No	Bachelor Degree level	Ongoing/Permanent (other than senior executive)	Female	$183,300 - $261,450	No	No
15	No	20 - 24	Less than 1 year	No	Graduate Diploma or Graduate Certificate level	Contract â€“ Non Executive	Female	$183,300 - $261,450	No	No
16	No	50 - 54	5 - 10 years	No	Graduate Diploma or Graduate Certificate level	Ongoing/Permanent (other than senior executive)	Female	$183,300 - $261,450	No	Yes
17	No	20 - 24	1 - 2 years	No	Prefer not to say	Ongoing/Permanent (other than senior executive)	Male	$121,917 - $140,995	No	No
18	No	40 - 44	More than 20 years	No	Prefer not to say	Ongoing/Permanent (other than senior executive)	Male	$121,917 - $140,995	No	No
19	No	45 - 49	10 - 20 years	No	HSC or equivalent	Ongoing/Permanent (other than senior executive)	Female	$183,300 - $261,450	No	Yes
20	Prefer not to say	20 - 24	10 - 20 years	No	Advanced Diploma or Diploma level	Ongoing/Permanent (other than senior executive)		$157,763 - $183,299	Prefer not to say	Prefer not to say
21	No	30 - 34	1 - 2 years	No	Bachelor Degree level	Ongoing/Permanent (other than senior executive)	Female	$183,300 - $261,450	No	No
22	No	40 - 44	More than 20 years	No	HSC or equivalent	Ongoing/Permanent (other than senior executive)	Male	$121,917 - $140,995	No	No
23	No	40 - 44	10 - 20 years	No	Advanced Diploma or Diploma level	Ongoing/Permanent (other than senior executive)	Female	$183,300 - $261,450	No	No
24	No	40 - 44	2 - 5 years	No	Bachelor Degree level	Temporary (including temporary teachers and graduates)	Female	$183,300 - $261,450	No	No
25	No	50 - 54	5 - 10 years	No	Less than year 12 or equivalent	Ongoing/Permanent (other than senior executive)	Female	$183,300 - $261,450	No	No

Figure 40 Sample of Synthetic NSW PMES Dataset

Dataset 9 – Synthetic NSW Workforce Profile Dataset (Synthetic Dataset)

Randomly generated dataset with fields including personal information (DOB, gender, country of birth, minority group status, highest education level, and disability status. Each individual synthetic government employee has a trajectory of changes in remuneration, legislation code, salary band, and standard weekly full-time hours over three years.

900,000 rows with 15 columns
300,000 unique synthetic employees (based on Gen_Code)

Gen_Code	DOB	Work and Live in Same Location Flag	Gender Code	Country of Birth	Disability Code	Highest Education Level Code	Language First Spoken Code	Minority Group Code	Year Workforce Profile	Salary_Band	Std_FT_Hours	Legislation_Code	Remuneration	Remuneration_Census
1	1964-04-2	N	2	-7777	2	-7777	1	2	2016	Clerk Grade 7 yr 2	38	35	91471.27432	3421.411834
2	1969-04-1	Y	1	-7777	-7777	-7777	1	2	2016	Clerk GS 6	31.25	20	44892.44343	2030.622414
3	1975-11-1	N	2	-7777	-7777	-7777	2	1	2016	Clerk Grade 8 yr 1	35	81	95481.22001	1899.34329
4	1954-05-2	N	2	-7777	4	-7777	1	2	2016	Clerk Grade 4 yr 2	38	401	75795.67887	3073.435251
5	1971-07-0	Y	2	-7777	4	-7777	1	2	2016	Clerk Grade 8 yr 1	35	81	94918.04368	3768.954304
6	1970-03-0	N	1	-7777	-7777	-7777	1	2	2016	Clerk Grade 4 yr 2	35	401	72353.07958	1542.115127
7	1975-06-1	N	1	-7777	4	-7777	-7777	-7777	2016	Clerk Grade 8 yr 1	49	81	94423.96834	475.0437418
8	1960-04-2	N	2	-7777	4	-7777	2	2	2016	Clerk GS 8	31.25	20	48009.80298	1426.972305
9	1986-01-0	N	1	Australia	4	-7777	1	2	2016	Clerk Grade 8 yr 1	49	81	95965.42005	3861.941059
10	1994-09-1	N	2	-7777	4	-7777	1	2	2016	Clerk GS 11	38	89	52847.64509	2022.843585
11	1959-06-2	N	1	-7777	4	-7777	1	2	2016	Clerk GS 13	38	35	56261.38806	2096.146678
12	1951-11-2	N	1	-7777	4	-7777	1	2	2016	Clerk GS 9	38	401	49708.73346	2115.735007
13	1974-09-2	N	1	-7777	4	-7777	1	2	2016	Clerk Grade 4 yr 2	38	65	71927.15094	2831.512033
14	1985-09-1	N	2	-7777	-7777	-7777	1	2	2016	Clerk Grade 10 yr 2	38	35	115808.7666	24693.53856
15	1980-02-2	N	1	-7777	-7777	-7777	1	2	2016	Clerk Grade 4 yr 2	35	81	74984.84252	3176.908184
16	1953-03-2	N	2	-7777	4	-7777	-7777	-7777	2016	> Clerk Grade 12 yr 2	35	402	158869.8456	15416.66761
17	1949-09-0	N	2	-7777	2	-7777	1	2	2016	Clerk Grade 4 yr 2	38	35	71973.95915	3072.80717
18	1986-01-1	N	1	-7777	4	-7777	-7777	-7777	2016	Clerk GS 8	31.25	20	48413.13612	940.9174792
19	1982-04-0	N	2	-7777	4	-7777	1	2	2016	Clerk Grade 2 yr 2	35	401	64316.19506	2429.769665
20	1974-01-0	N	2	-7777	-7777	-7777	1	2	2016	Clerk Grade 5 yr 2	49	81	82022.75312	3272.934917
21	1987-02-1	N	2	-7777	4	-7777	1	-7777	2016	Clerk Grade 6 yr 2	38	65	85653.22922	3557.38843
22	1981-06-2	N	2	-7777	-7777	-7777	-7777	-7777	2016	Clerk Grade 8 yr 1	35	81	95188.16617	2838.776711
23	1978-06-2	N	2	-7777	4	-7777	1	2	2016	Clerk Grade 11 yr 1	35	402	117897.894	4426.17807
24	1955-08-0	N	2	-7777	4	-7777	1	2	2016	Clerk Grade 8 yr 1	49	81	96460.49633	1717.604928

Figure 41 Sample of Synthetic NSW Workforce Profile Dataset