December 2019

Privacy-Preserving Data Sharing Frameworks

People, Projects, Data and Output

About the editor

Dr Ian Oppermann, FACS CP

ACS VICE PRESIDENT – ACADEMIC BOARDS

Ian has over 20 years' experience in the ICT sector and has led organisations with more than 300 people, delivering products and outcomes that have impacted hundreds of millions of people globally. He has held senior management roles in Europe and Australia as Director for Radio Access Performance at Nokia, Global Head of Sales Partnering (network software) at Nokia Siemens Networks and Divisional Chief and Flagship Director at CSIRO. Ian is considered a thought leader in the area of the digital economy and is a regular speaker on big data, broadband-enabled services and the impact of technology on society. Ian has an MBA from the University of London and a Doctor of Philosophy in Mobile Telecommunications from the University of Sydney.

Many people and groups came together to make this report a reality. We'd like to give thanks to the following organisations for their assistance in putting together this report.

Foreword

It's an unavoidable fact of our modern world that massive amounts of data is being collected about every one of us. Every single day, there are literally thousands of data points about you that are being collected, stored and analysed.

It's not just your online activities. You're generating data every time you go past a traffic monitoring camera; when you get on and off the train; when you walk past a security camera; or turn on an appliance, affecting your power usage; when you watch pay TV; or use your GPS and a million other activities besides. You're generating data. That data is often shared between organisations, or used for data analysis.

That's not inherently a bad thing. It's this data that's enabling the smart services that are making our world more liveable. Data gathering and sharing between organisations is allowing governments to make better and faster decisions about resource allocation. It's allowing companies to better target the needs of their customers.

The danger is when that data gets abused in ways that destroy our privacy and right to anonymity. Many of the data points mentioned above are ostensibly anonymous, but smart attackers can link data points together to identify and gain information about individuals. De-identified shared datasets can be re-identified using similar linking techniques.

The challenge for data scientists, for government policy makers, for businesses that gather and use data about individuals (which is most businesses) is how to enable all the benefits of those smart services and data sharing without taking away the fundamental right to privacy of our citizens.

For three years ACS has worked with many of Australia's leading data scientists and companies to develop a workable, practical solution to this problem. This paper is the third in a series of papers that are working towards that invaluable goal: how to preserve privacy while enabling smarter services through shared data.

Solving this in a practical, workable way will provide massive benefits to Australian companies and organisations, allowing them to more effectively access and offer shared datasets without the fear of tripping over privacy laws.

This paper focuses on the practical aspects of de-identifying data and applying the Five Safes framework, and is the next milestone in solving this problem. I'd like to thank everybody who has worked on it and continues to work to ensure that Australia can maintain its fundamental rights to privacy while gaining all the benefits of modern data sharing.

The Hon Victor Dominello MP
NSW Minister for Customer Service

Foreword by ACS

In the 21st century it can almost seem like privacy is a thing of the past, a quaint notion from the era before big data became a thing. Hardly a week goes by now without a new privacy scandal making the headlines. Social media, mobiles, smart devices, open data, ubiquitous connectivity – all these things are making it ever harder to maintain the essential right to privacy of every person.

Yohan Ramasundara
President, ACS

Andrew Johnson
Chief Executive Officer, ACS

The billion dollar question is: what can we do about it? Is it even possible to enjoy those smart services without sacrificing at least some of our privacy?

The answer, we believe, is yes. But it's unbelievably hard, and that's why ACS has been working with Australia's leading experts for more than three years now to crack one of the key pillars of privacy in the 21st century: how to preserve privacy in shared datasets.

For Australia and the world, data is a key strategic asset. It's a huge driver of productivity growth and has the potential to unlock incredible value for Australia's industries. In 2016, the Commonwealth government predicted that open government data alone would be worth $25 billion to the Australian economy, as well as provide innumerable value to researchers, NGOs and other organisations.

And that's just open government data. Other organisations – researchers, businesses, educational institutions – can add value by providing and using shared datasets.

The trick, as it were, is how to do that without giving up the personal information of the people whose data is contained in those datasets. The ethical and moral boundaries of data sharing are subjects that have been consuming ACS and its members for years, and why we've worked so hard through these series of white papers to answer some of the subject's most glaring issues.

A great number of Australia's leading data scientists have contributed valuable time and insights to these works, and we'd like to thank them all. Led by ACS Vice President (Academic Boards), the inestimable Dr Ian Oppermann, these experts have produced world-first works on the subject of practically measuring and quantifying the presence of personal information in a dataset. Nowhere else in the world has this been done, and it's incredible that a team of volunteer researchers in Australia has not only taken on this challenge, but has risen to meet it so spectacularly.

Contents

Executive summary

This paper describes a framework for privacy-preserving data sharing, addressing technical challenges as well as data sharing issues more broadly. It builds on the 2018 ACS Report, *Privacy in Data Sharing: A Guide for Business and Government*, expanding the concept of a Personal Information Factor and introducing a Utility Factor with worked examples. It describes frameworks for data sharing that consider both the Personal Information Factor in the data, as well as sensitivities in the data and sensitivities in use of outputs of analysis of data.

This work in this paper has been underpinned by a series of ACS Directed Ideation technical workshops which focused on considerations and controls for data use. The conclusions presented in this paper have been informed by this workshop series.

CONCLUSION 1

Many of the voiced concerns about data sharing are expressed as concerns about privacy. In practice they are based on concerns about the sensitivity of data and use of outputs.

CONCLUSION 2

The use case for data strongly influences the risk framework required and the methods (aggregation, suppression, obfuscation, perturbation) appropriate for increasing data safety.

CONCLUSION 3

It is feasible to develop a meaningful Personal Information Factor (PIF), giving a measure of personal information in de-identified, people-centric data. Information theoretic metrics show promise for many common protection methods and can be enhanced to cover perturbed data.

CONCLUSION 4

Re-identification risk and levels of personal information in data are related but different concepts. Additional work is needed to relate the re-identification risk metric to the legal definition of privacy, and the assumed attacker model.

CONCLUSION 5

Understanding the relationship between different features in a dataset helps to identify those features that carry the highest information and risk of re-identification, as well as those that have the greatest impact on data utility after protection methods are applied.

CONCLUSION 6

Development of a meaningful measure of relative utility is feasible for datasets protected through aggregation, generalisation, obfuscation and perturbation. Information theoretic metrics based on mutual information (between original and protected datasets) shows promise.

CONCLUSION 7

Dealing with "trajectories" (or pathways) in data is critical to its safe use and release. Developing methods to address trajectories is possible. The methods explored in this paper show promise; however, the complexity of the approaches may limit real-world implementation.

01

Introduction

The problem

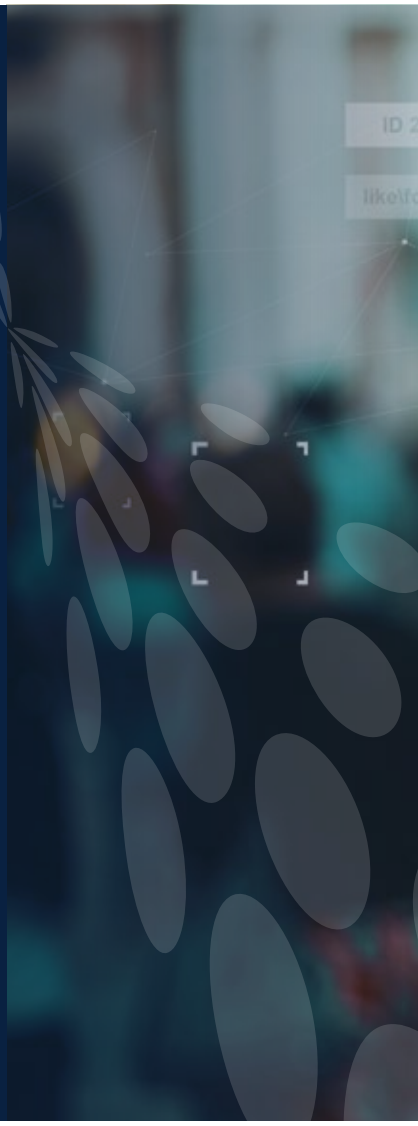
Future smart services for homes, factories, cities and governments rely on sharing of large volumes of often personal data between individuals and organisations, or between individuals and governments.

A smart light in your home that turns on and off as you move around the house can provide efficient use of energy for lighting, but will develop de-identified data about when you are home, which rooms you use and when, if there are other people in your home, and where in your home you spend your time.

Within this de-identified data there are insights about you, your relationships, habits and preferences. In aggregate form, this data can be used by a smart lighting provider to deliver more efficient lighting services to a suburb, by a smart grid to match energy demand to energy supply, or by a smart micro energy service provider to make best use of spot energy prices.

And that's just one of the many smart services that are possible and available right now. Consider how many "smart" devices and services we use in our personal lives (smart TV, smart scales, smart toilet, smart phone, virtual assistant, smart home, smart car), or in the wider community (smart grid, smart materials, smart factory, smart city, even smart government). The benefits of these "smart" services in terms of improved efficiency, improved effectiveness and increased personalisation to our individual needs is enormous.

However, there's a major problem that need to be addressed. If these datasets are linked, a great deal of personal



THE CHALLENGE IS TO ADDRESS THE

BIG ISSUES AROUND SHARING OF DE-IDENTIFIED DATA,

BROADLY ARTICULATED AS THE NEED TO

PROTECT INDIVIDUAL PRIVACY

AND CONCERNS ABOUT UNINTENDED CONSEQUENCES OF DATA SHARING.

information may be contained in the joined data, sufficient to reasonably re-identify individuals represented in the data. How this data is used and by whom creates risks and concerns.

The challenge is to address the big issues around sharing of de-identified data, broadly articulated as the need to protect individual privacy and concerns about unintended consequences of data sharing.

Over the last three years, ACS has been exploring these aspects of data sharing in an attempt to develop useful risk frameworks that address these concerns and, where possible, quantify protections to these risk frameworks.

The focus for its most recent activities has been privacy and personal information. Much of the basis for the risk frameworks presented in this paper are built on the foundational assumption that we can develop a measure for the amount of personal information in a linked de-identified dataset. This measure – what we call a Personal Information Factor (PIF) – has developed to become a measure of uniqueness of the members of a dataset and the information gained from identifying these individuals.

Throughout this paper, a number of datasets have been used to create example results for illustrative purposes and to test algorithms developed throughout the work underpinning this paper.

The datasets are described in Appendix A of this paper and include a number of open datasets from the US and Australia, synthetic datasets created from actual NSW Government data, and one randomly generated set using statistical shaping based on known distributions (hospital admissions).

This paper assumes all analysis is performed using de-identified data. It is also assumed that the de-identified data is not subject to any national security classification.

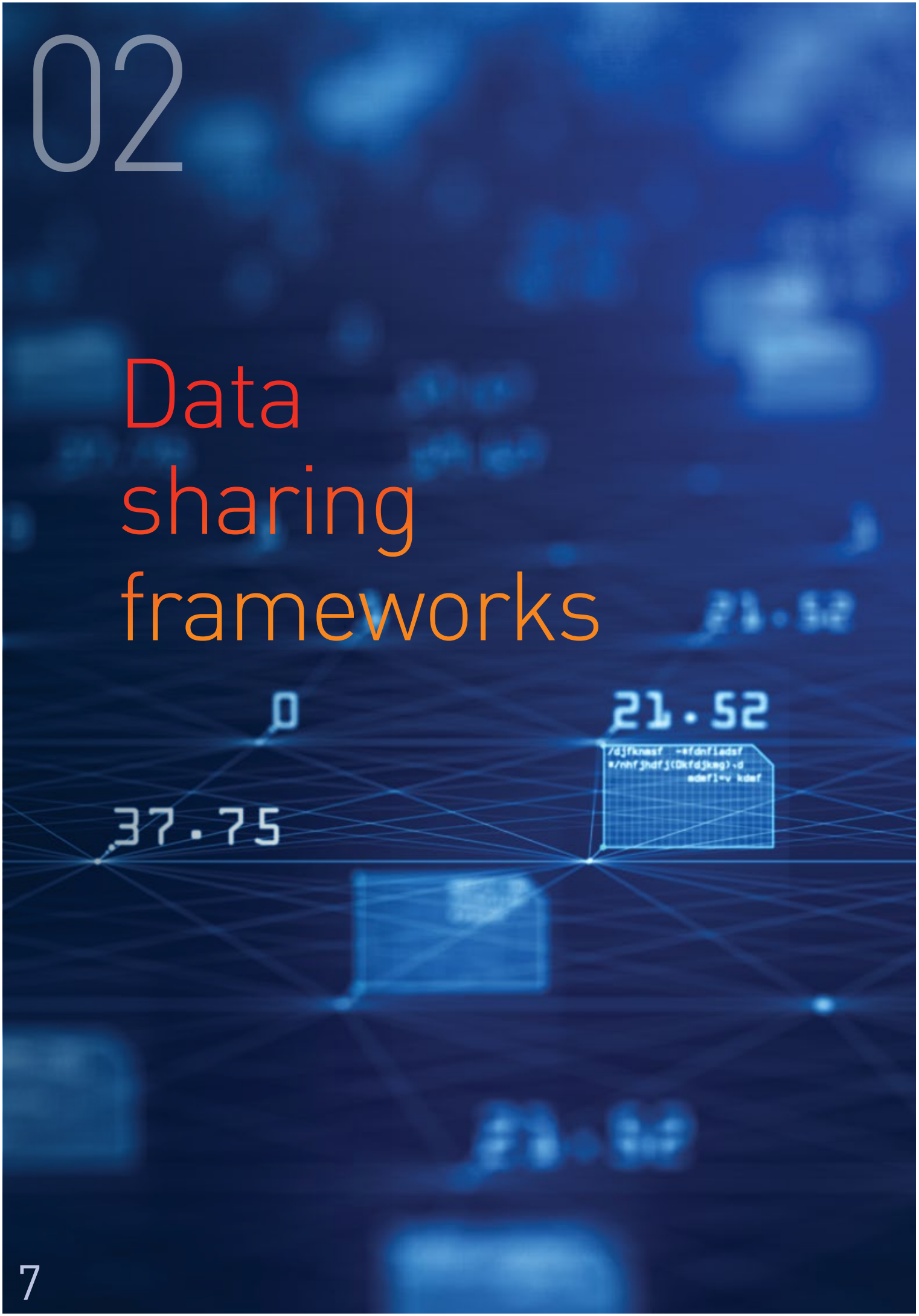
Differentiating between information and data

The terms "information" and "data" are often used interchangeably, reflecting the real-world human process of interpretation. However, for the purposes of this paper, *information* refers to an insight gained from analysis of data whether processed by an algorithm or a human being. An algorithm may process a year of financial data and identify potential fraud. A human being may look at dataset of patient admissions for rare diseases and recognise someone they know (or believe they know).

In both cases, there is processing of the data. In the case of the human being, there is also personal context and knowledge of the world added to the cognitive processing. For the purposes of this paper, data is the recoded sampling of the world or system or process. Data that records rare or unexpected events is said to have high information content. Data that records entirely expected or very common events is said to have low information content. It is the processing of this data that reveals the *information* contained within.

02

Data sharing frameworks



Considerations for data sharing – privacy, sensitivity, consequences and harm

The intended use of the data is a very significant factor when determining the risk framework for data sharing and use. Most often a use case is described in terms of:

- Who wants to access the data.
- Why they want to access the data.
- Consideration of the level of personal information in the data.
- Consideration of aspects of sensitivity of the data and the results of analysis.
- Concerns about the level of granularity of access to the data.
- Concerns related to the use of insights and decisions generated from analysing the data.

Without considering aspects of privacy, sensitivity or regulatory environment, the general use cases for data sharing and use can be stated as an interaction between a "Holder" and "User" of data:

- **Use case 1:** Holder shares actual data or data products with User.
- **Use case 2:** User can query data and gain insights but not directly access data (the "vault" model).
- **Use case 3:** Holder shares actual data or data products with User, and User can modify data.

The challenge for a use case is to determine which dimensions are set by the nature of the problem and which need to be adjusted in response to the nature of the problem. Aspects of sensitivity of use of data can include:

- Sensitive subjects recorded in data (subjective but often described in different economies such as data on health, religion and sexual orientation).
- Concerns about unexpected insights being generated from data, leading to negative surprises or embarrassment of data holder.
- Concerns about who can see or use insights generated from data.
- Concerns about the data user's ability to appropriately interpret results (the belief that expert knowledge or context is required).
- Concerns about insights or decisions from data (outputs) generated from poor-quality data.
- Concerns about insights or decisions from data (outputs) generated from poor-quality analysis.

- Concerns about unintended consequences of insights or decisions from data (outputs) being used.
- Concerns of loss of agency (control) for the data holder.
- Concerns about the age of data (previously unexamined data; data that describes contemporary situations; or data that was gathered in an environment which is no longer current, meaning outputs require contextualisation).
- Concerns about accidental release of data or insights (outputs).
- Concerns about being able to explain the action made based on an insight or decision from an analytical output.
- Concerns about the reversibility (or not) of an action taken based on an insight or decision from an analytical output.
- Concerns about harm caused based on an insight or decision from an analytical output.

A REAL-WORLD EXAMPLE OF USE CASES:

Lake temperature

A local council wants to develop of a water temperature heatmap for an environmentally sensitive lake. Data measurement will be through a network of water-based sensors that are sparsely spread and many of which are located near isolated lakeshore homes. The data therefore has the potential to reveal information about occupancy of the homes or activities taking place within the homes. Some of the basic aspects to consider are:

- **Project (fixed)** — the merits of the project may well provide a strong motivation to proceed.
- **Data (fixed)** — the location of sensors near isolated homes means that the data is highly likely to contain personal information.
- **People (variable)** — a high likelihood of personal information in data means protections must be put in place to limit the people who access the data or carry out the project.
- **Setting (variable)** — a high likelihood of personal information in data means protections need to be put in place to limit access to data and outputs of analysis.
- **Outputs (variable)** — the project requires only aggregated output so the results of analysis can be treated to reduce the level of personal information before release.

In this example, the high-level output may be aggregated in a temporal or spatial sense to reduce the re-identification risk and reduce the amount of information released.

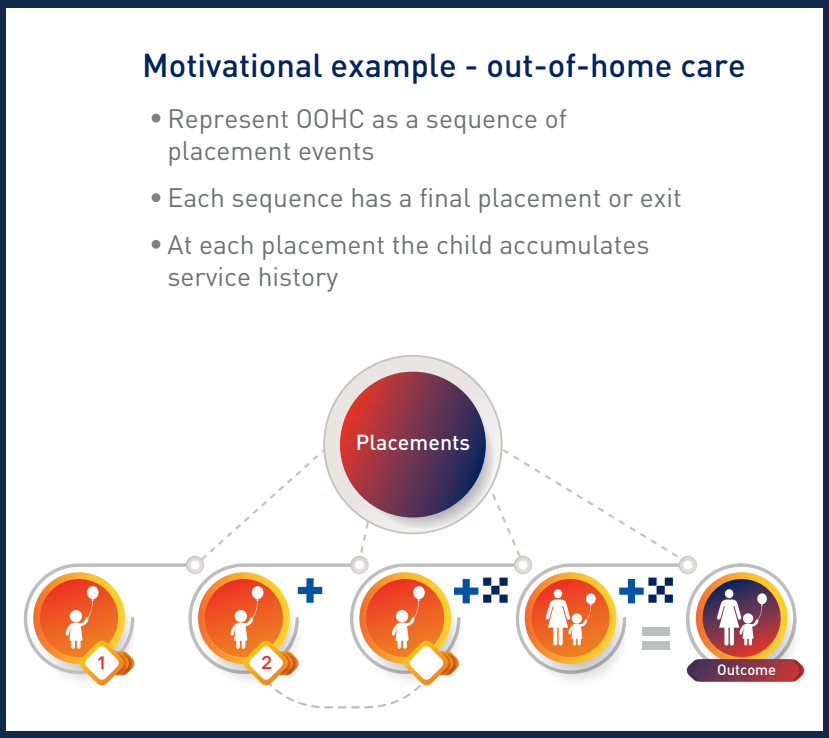


FIGURE 1. MOTIVATIONAL EXAMPLE, OOHC REFORM

Reforming out-of-home care

As another real-world example, consider the reform of out-of-home care (OOHC) in NSW, Australia1 (see Figure 1).
The OOHC scheme works with children who have been identified as being at risk of significant harm, placing such children into protective environments.

The scheme's reform is underpinned by the creation of longitudinal datasets, linking data from many government agencies on an individual (child-centric) basis.[1] All data is de-identified before linkage.

Nonetheless, concerns persist about privacy and sensitivity about the use of data, the nature of the project and use of outputs.

In the OOHC example, concerns identified included:

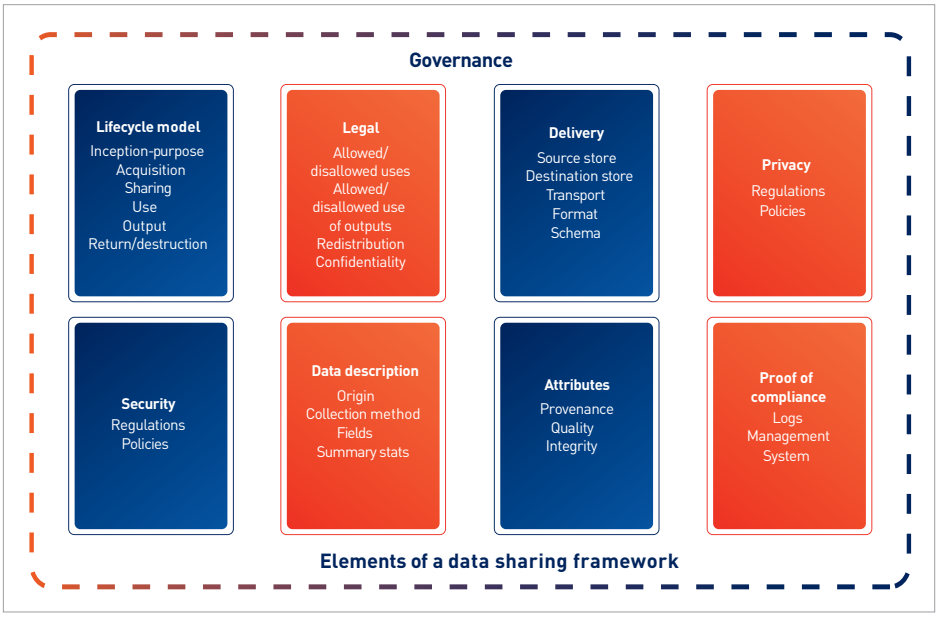
- What if a machine or algorithm generates insights (outputs)? Can the results be trusted?
- Who can access these datasets?
- Who are outputs shared with?
- What are the consequences of sharing or using these insights (outputs)? Can this make things worse (outcomes)?
- What if linked de-identified data contains sufficient personal information to reasonably identify individuals?
- Will poor data quality lead to inaccurate insights?
- How do we ensure that there is always a human in the loop, so that a machine or an algorithm is not empowered to automatically act on the insights generated?
- How do we ensure appropriate access and authorisation to data and analytical insights?

[1] For more details of data assets used for reform, see https://www.theirfuturesmatter.nsw.gov.au/

Governance dimensions for data sharing frameworks

Issues of data sharing and use are acknowledged to exist throughout the lifecycle of data creation, collection, storage, use, analysis, reuse, archiving and deletion. If the inherent "risk" of data use increases during the lifecycle of a project, then the protections needed must also increase over time to ensure the project remains "Safe". The aspects of governance that needs to be considered at different stages of the lifecycle are shown in Figure 2. How these may be used at different phases of a project is shown in Figure 3. In this diagram, PIA refers to a Privacy Impact Assessment, a point-in-time evaluation of the potential impact on privacy.[2]

FIGURE 2. GOVERNANCE ASPECTS TO BE CONSIDERED FOR DATA USE AT DIFFERENT PHASES OF THE DATA LIFECYCLE.



[2] See, for example, the Office of the Australian Information Commissioner. Available online at https://www.oaic.gov.au/privacy/guidance-and-advice/guide-to-undertaking-privacy-impact-assessments/#introduction-to-privacy-impact-assessments

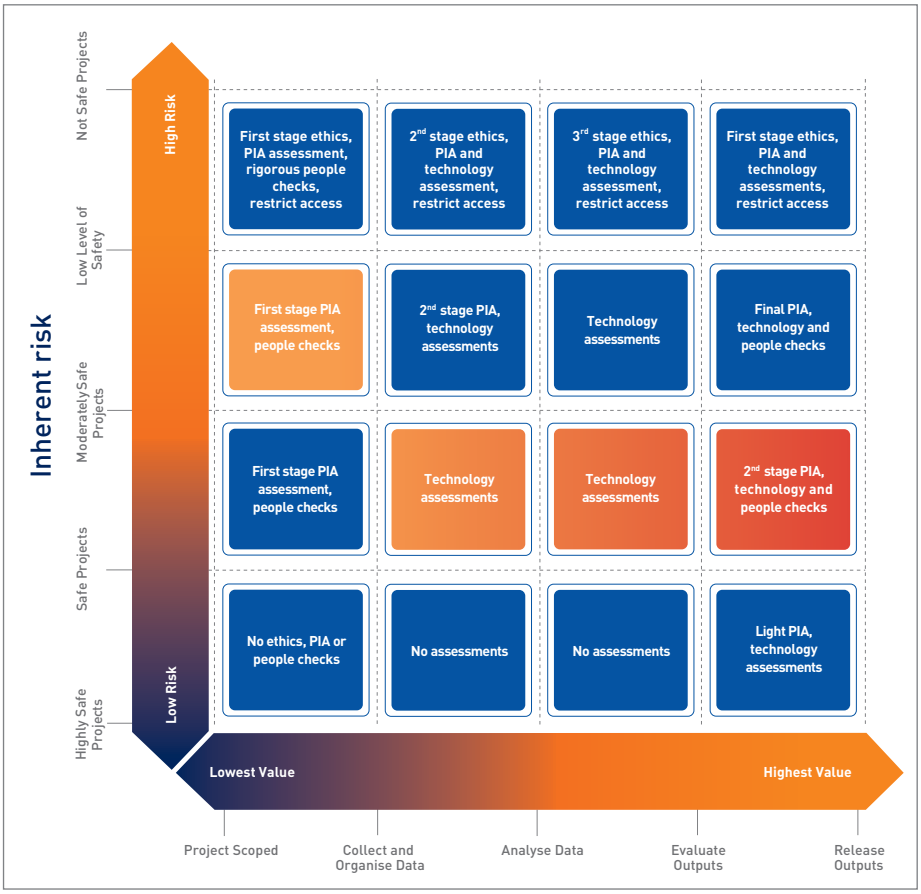


FIGURE 3. DATA LIFECYCLE AND EXAMPLE GOVERNANCE AT STAGES OF THE LIFECYCLE

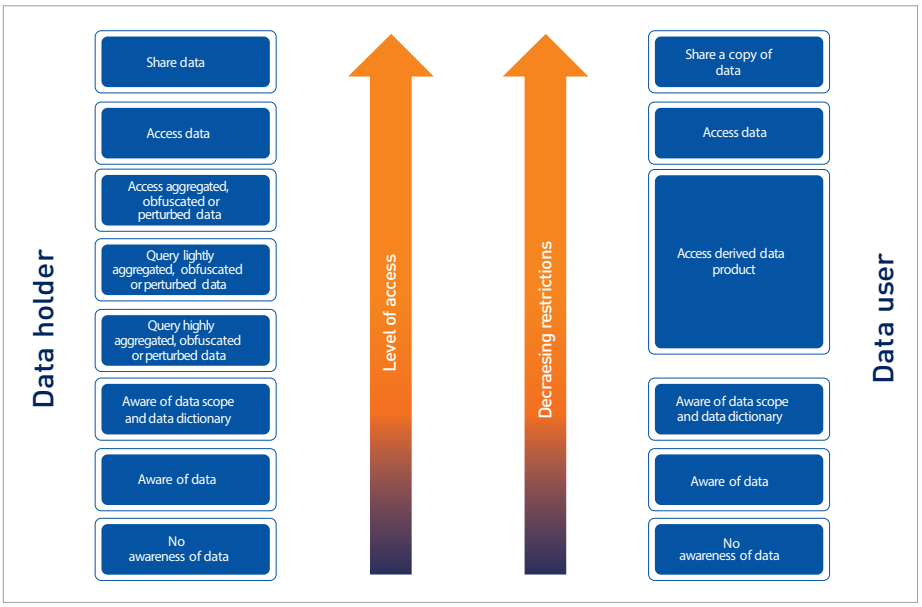
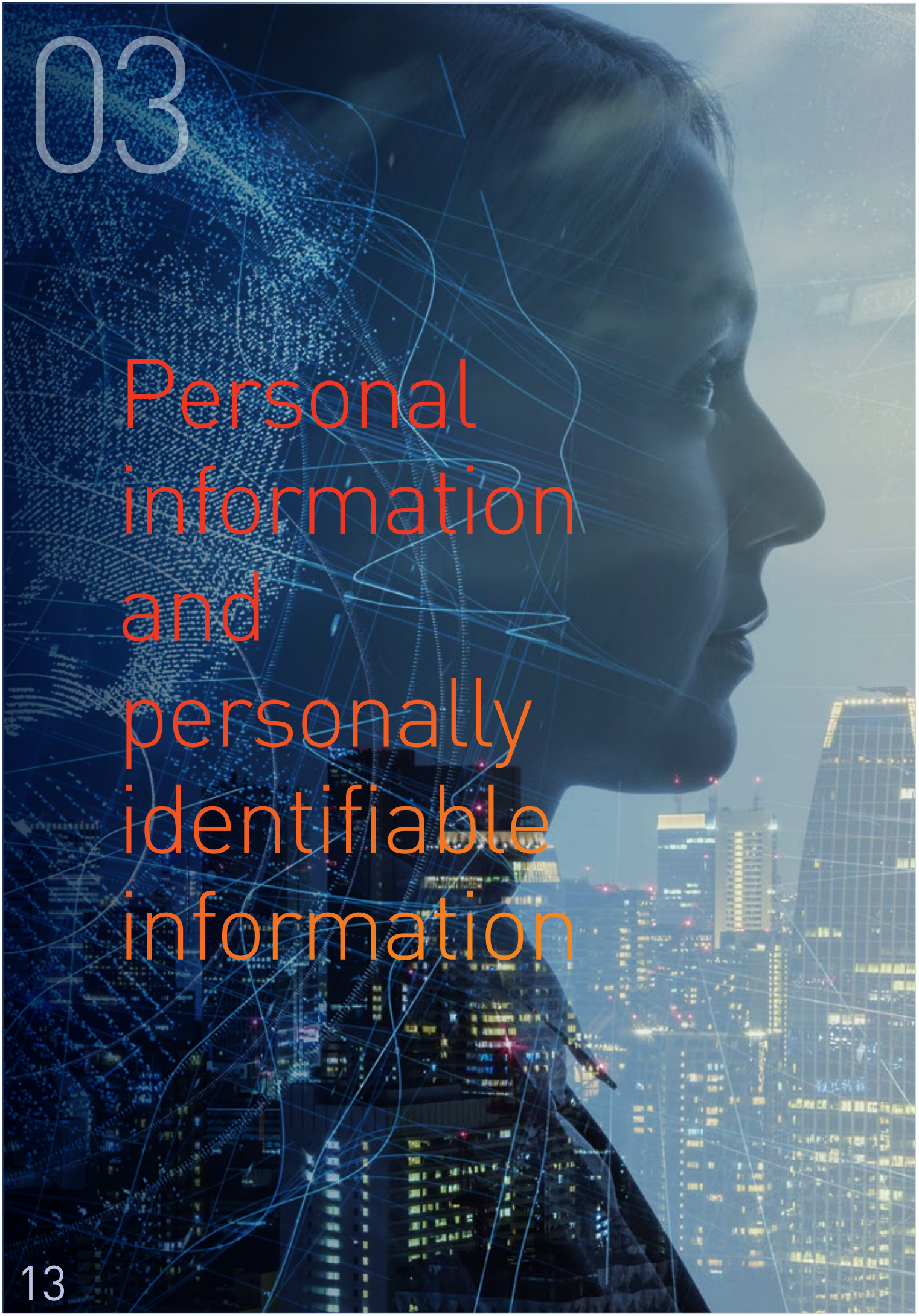


FIGURE 4. FRAMEWORK FOR DATA SHARING AND USE

A general framework for data sharing and use is shown in Figure 4. This diagram highlights sharing from the most restrictive to least restrictive. The restrictions range from not sharing knowledge that the dataset exists, to allowing access to data products, through to sharing datasets. The "Data Products" described are intended to include metadata, provenance data (a specific form of metadata), aggregated data and modified versions of the underlying data.The different levels of access provide opportunities for different controls for risks identified during different phases of a project.



03

Personal information and personally identifiable information

Many of the concerns about collection, use and sharing of data relate to privacy. As a consequence, in many parts of the world there are legislative frameworks that require very strict controls to be placed around the use of personal information.

The terms personal information (PI) and personally identifiable information (PII) are often used interchangeably in different legislative frameworks around the world. Date of birth is often considered personal information (ie. information about a person), but used alone, this single feature is not personally identifiable information unless it uniquely identifies an individual. The question becomes: how many features are needed to be linked before personal information becomes personally identifiable information for a person known to be in a dataset?

A recent paper published in *Nature Communications*[3] provides a means to estimate the likelihood of a specific person to be correctly re-identified, even in a heavily incomplete dataset. This paper is the latest in a long series of works that show a small number of features can be linked to identify a unique individual in a population. The "heavily incomplete" aspect of the paper shows the limitations of the protection associated with creating uncertainty as to whether an individual is in a dataset.

The concepts of personal information versus personally identifiable information are not clearly differentiated in regulatory frameworks. Personal information is typically described so as to cover a very wide field and is described differently in different parts of the world. For example, in the state of 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.

A major focus of this paper is the attempt to build quantified measures and risk frameworks for "reasonably" in different contextual environments.

"PERSONAL INFORMATION (INCLUDING INFORMATION OR AN OPINION FORMING PART OF A DATABASE AND WHETHER OR NOT RECORDED IN A MATERIAL FORM)) MEANS INFORMATION OR AN OPINION ABOUT AN INDIVIDUAL WHOSE IDENTITY IS APPARENT OR CAN REASONABLY BE ASCERTAINED FROM THE INFORMATION OR OPINION"

[3] L. Rocher, J. M. Hendrickx and 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

Separating sensitivity and personal information

While privacy is often cited as the main reason to restrict to data sharing, there are other concerns that relate to unintended consequences of the use of data including:

- The release of data about vulnerable individuals.
- The loss of exclusive access to insights from data.
- The appropriate use of the data.
- The appropriate use of insights gained from data.
- Unexpected or embarrassing results found from analysis of data.
- A recipient's lack of expertise to analyse or interpret the data.
- The data's age and quality.
- The use of data without the context knowledge of its collection or data quality.

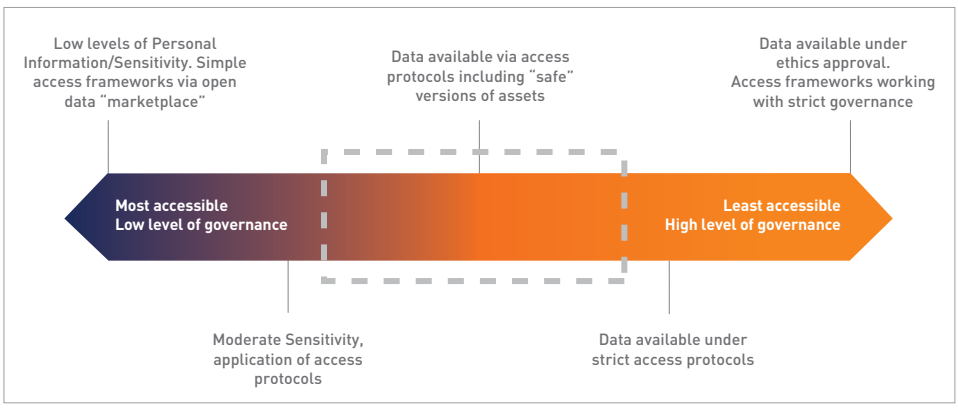


FIGURE 5. CONCEPTUAL SENSITIVITY SCALE

Broadly, these are described as sensitivities of data and will be explored separately from privacy concerns. Figure 5 highlights the need for lower (left-hand side) or higher (right-hand side) levels of governance, support or expert interpretation required for use of data (and production of outputs) of different sensitivity. Figure 6 shows a conceptual framework which allows us to consider the aspects of sensitivity separately to privacy.

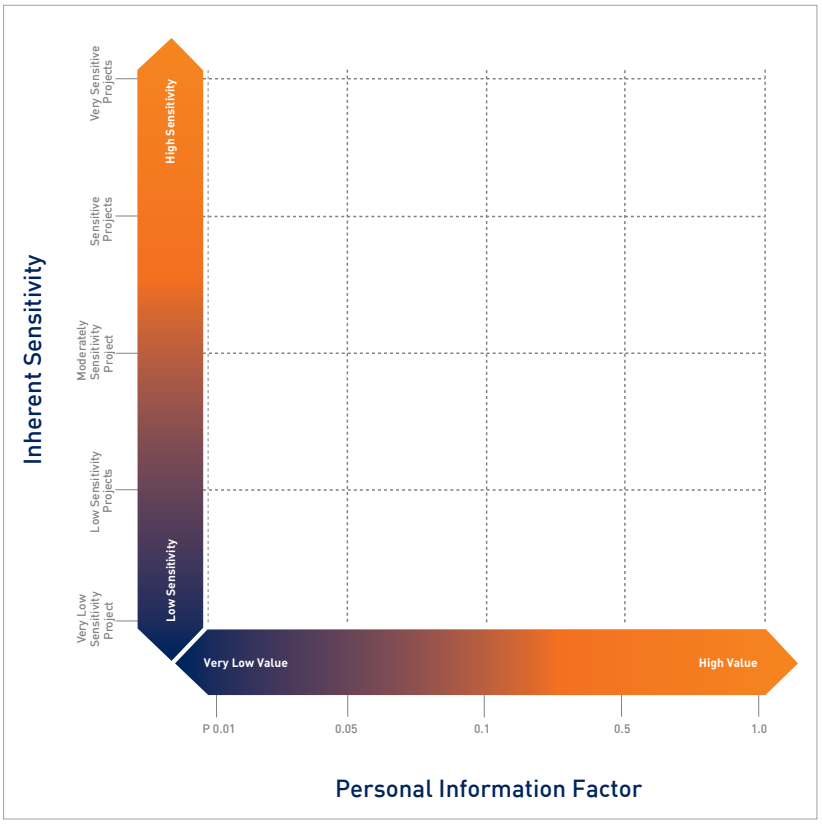


FIGURE 6. SENSITIVITY VERSUS PRIVACY

A multidimensional risk framework – Five Safes

In October 2017, the Australian Computer Society (ACS) released a technical white paper that explored the challenges of data sharing[4]. 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[5] for data sharing[6]. The modified framework attempts to quantify different thresholds for "Safe". In November 2018, ACS released a second technical white paper on *Privacy in Data Sharing*[4], which evolved the concepts introduced in the first paper.

The white papers 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 that is confidential or sensitive. The dimensions of the framework are shown on the right.

[4] See ACS website, available online https://www.acs.org.au/content/dam/acs acs-publications/ACS_Data-Sharing-Frameworks_FINAL_FA_SINGLE_LR.pdf

[5] T. Desai, F. Ritchie, R. Welpton, "Five Safes: Designing data access for research", October 2016, https://uwe-repository.worktribe.com/output/914745

[6] See ACS website, available online https://www.acs.org.au/content/dam/acs/acs-publications/Privacy%20in%20Data%20Sharing%20-%20final%20version.pdf

SAFE PEOPLE

Refers to the knowledge, skills and incentives of the users to store and use the data appropriately. In this context, "appropriately" means "in accordance with the required standards of behaviour", rather than level of statistical skill. In practice, a basic technical ability is often necessary to understand training or restrictions and avoid inadvertent breaches of confidentiality; an inability to analyse data may lead to frustration and increases incentives to share access with unauthorised people.

SAFE PROJECTS

Refers to the legal, moral and ethical considerations surrounding use of the data. This is often specified in regulations or legislation, typically allowing but limiting data use to some form of valid statistical purpose, and with appropriate public benefit. Grey areas might exist when exploitation of data may be acceptable if an overall public good is realised.

SAFE SETTING

Refers to the practical controls on data access. At one extreme, researchers may be restricted to using the data in a supervised physical location. At the other extreme, there are no restrictions on data downloaded from the internet. Safe settings encompass both the physical environment (such as network access) and also procedural arrangements such as supervision and auditing regimes.

SAFE DATA

Refers primarily to the potential for identification in the data. It may also refer to the quality of the data and the conditions under which it was collected, the quality of the data (accuracy), the percentage of a population covered (completeness), the number of features included in the data (richness), or the sensitivity of the data.

SAFE OUTPUTS

Refers to the residual risk in publications derived from sensitive data.

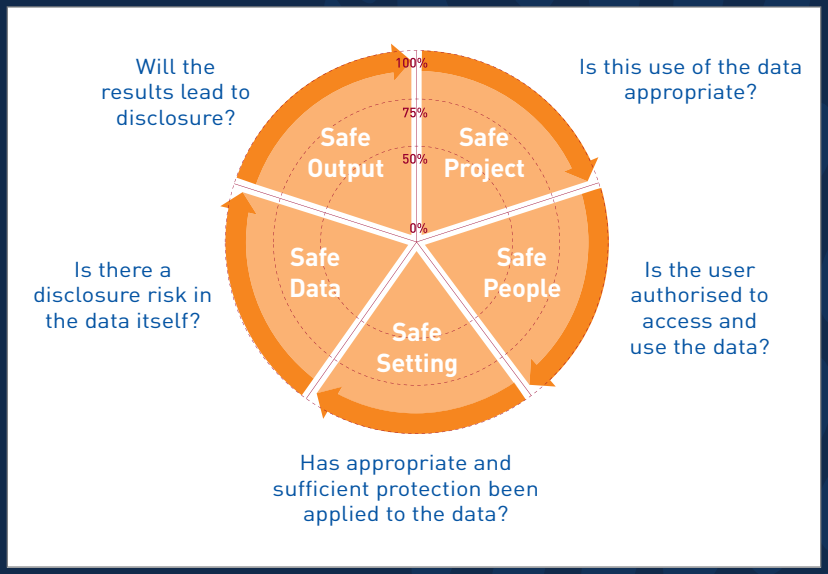


FIGURE 7. MODIFIED FIVE SAFES FRAMEWORK

The Five Safes framework is relatively easy to conceptualise when considering cases of "extremely Safe" data, although it does not unambiguously define what this is. An extremely Safe environment may involve researchers who have had background checks, projects that have ethics approval and rigorous vetting of outputs from that data environment. Best practice may be established for such frameworks, but none of these measures is possible to describe in unambiguous terms as they all involve judgement.

The adapted model explores different, quantifiable levels of "Safe" for each of People, Projects, Setting, Data and Outputs, as well as how these different "Safe" levels could interact in different situations. Figure 7 shows the dimensions of the adapted "Five Safes" framework taken from the 2018 ACS technical white paper.

One of the great challenges of this model is the interaction between the risk dimensions. The Project or purpose can impact People, Data, Settings and Output; Data can impact People, Setting and Output. The ability to work out which risk dimensions are fixed, and which need to be adapted in response to these risk frameworks makes the approach an iterative process at best.

A higher-order risk framework – a few "Safes" more

As the work of the ACS Data Sharing Taskforce has continued, it became clear that the scope of the five dimensions of risk were insufficient to address the risks of use of data for analysis or for data sharing. Additional risk dimensions were identified including the following.

SAFE ORGANISATION

Refers to the systems and processes employed by an organisation to ensure the Five Safes framework is applied throughout the project and with the long-term management of data and outputs. Safe organisations may include those that adhere to data protection, quality standards and cyber security standards. Safe Organisation may consider:

- Quality control systems and processes.
- Ethical or appropriateness checks of individual projects.
- Security and quality screening processes for People.
- Data governance systems and processes.
- Output release systems and processes.
- Technical expert output review process.
- Transparent process review.
- Cyber physical security.

SAFE LIFECYCLE

Refers to the time sensitivity of a dataset or output. Data may be highly sensitive for a specific period and then may be not sensitive at all. For example, a city plan that might involve the mandated acquisition of an individual's home to enable the construction of a new road may be very sensitive until the home is demolished. At this time there is no remaining value in protecting the data or output. Considering the complete lifecycle of a dataset may add additional insight and tools to help effectively anonymise and protect privacy rights. Safe Lifecycle may consider:

- Data archive and governance systems and processes maintained over time.
- Output archive and governance systems and processes maintained over time.
- Controls on People maintained over time.
- Cyber physical security maintained over time.

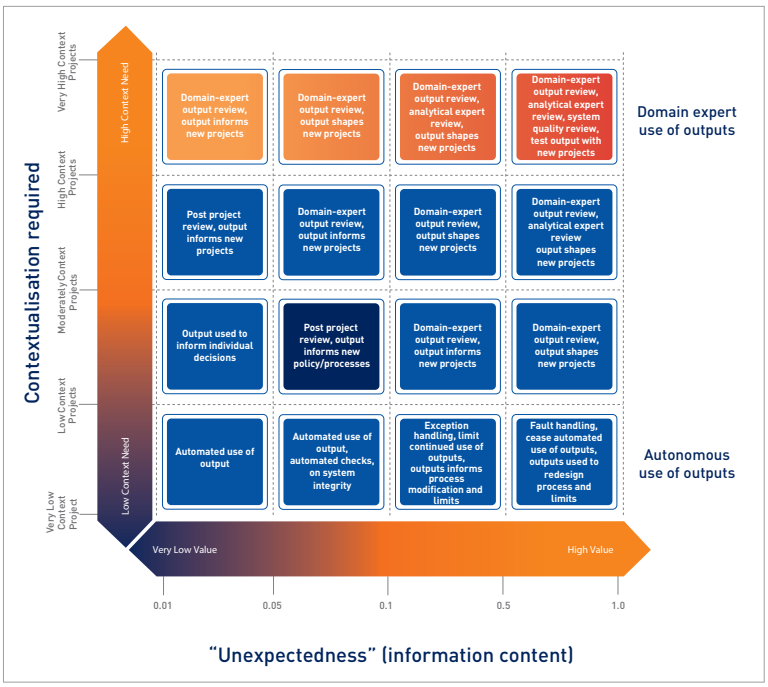


FIGURE 8. SAFE USE - GOVERNANCE FRAMEWORKS ADDRESSING TWO FACTORS OF "SENSITIVITY"

Figure 8 illustrates a possible framework for "Safe Use" of data and Outputs when considering two dimensions of Sensitivity – unexpectedness of result and the level of "expert" interpretation or contextualisation required. We'll cover this more in Chapter 7.

SAFE OUTCOMES

Refers to the ultimate uses of the project outputs. A variety of Outcomes Frameworks have been developed that can be informed by the outputs of individual data linkage and analysis projects. Safe Outcomes may consider:

- Transparency of Outcome goals.
- Inclusive development of Outcome goals.
- Appropriateness review of work programs and individual projects.
- Domain expert output review processes.
- Output use governance and processes.
- Review of impact of use of outputs.
- Transparent process review.
- Cyber physical security.

SAFE USE

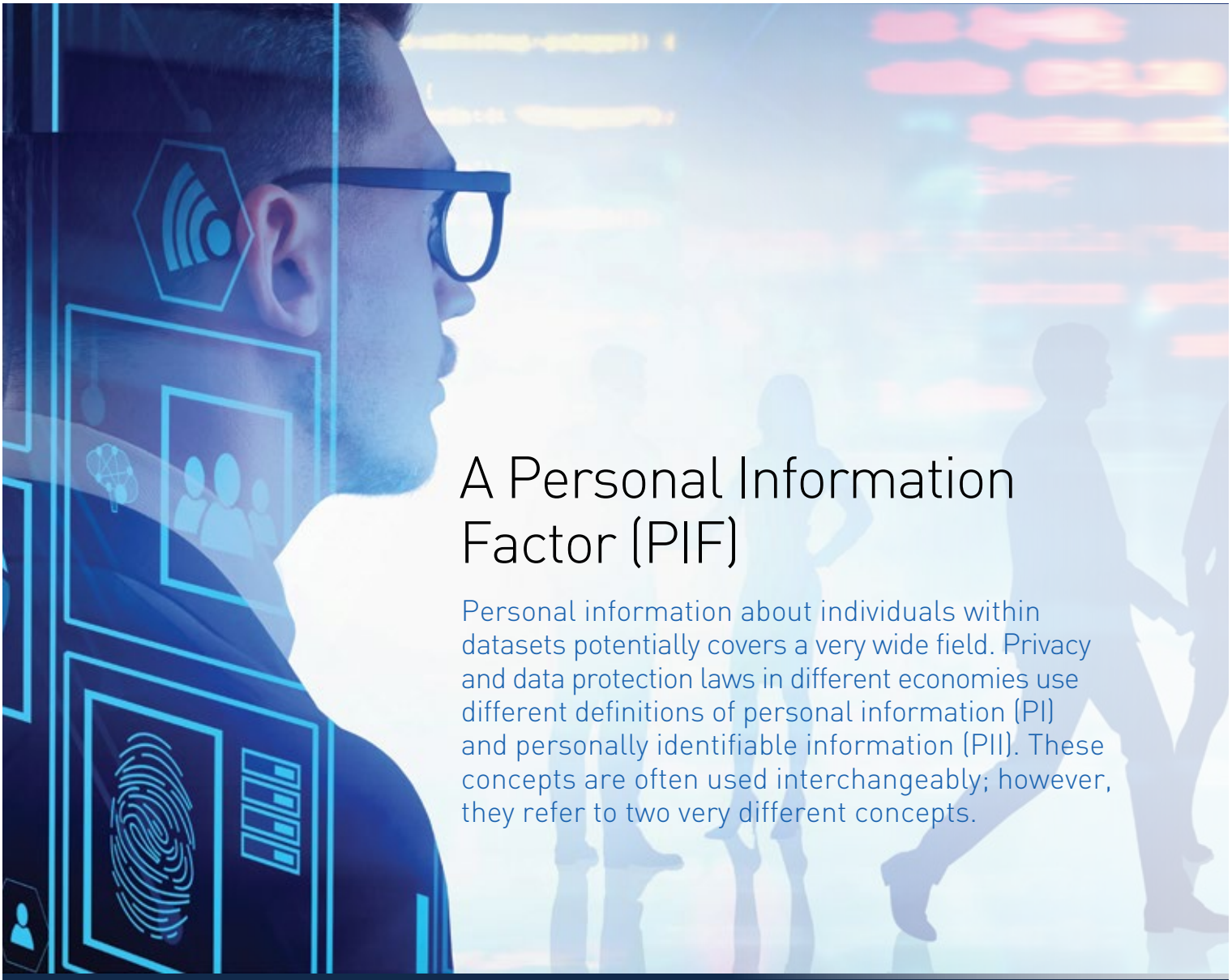
Refers to the use of the outputs within the Outcomes Framework specifically. This includes how much interpretation or context is required to appropriately use the outputs, including the degree to which a decision or action can be informed or automated based on this output. Safe Use may consider:

- Frameworks to evaluate consequence of use of output.
- Frameworks to determine the level of confidence in outputs accuracy (quality).
- Frameworks to determine the completeness of outputs (how much contextualisation is needed).
- The degree of automation associated with the use of output.
- Frameworks to evaluate the "expectedness" of outputs (high or low information content).
- Long-term monitoring of consequences of use of output.

SAFE RESPONSE

Refers to the systems and process which need to be in place to address adverse consequences of sharing of data or actions taken based on outputs. Safe Response may consider:

- Systems and processes to identify accidental release of data or outputs.
- Systems and processes to respond to accidental release of data or outputs.
- Systems and processes to communicate accidental release of data or outputs.
- Systems and processes to assess impact of accidental release of data or outputs.
- Systems and processes to support real-world compensation in response to accidental release of data or outputs.



A Personal Information Factor (PIF)

Personal information about individuals within datasets potentially covers a very wide field. Privacy and data protection laws in different economies use different definitions of personal information (PI) and personally identifiable information (PII). These concepts are often used interchangeably; however, they refer to two very different concepts.

Previous ACS technical white papers explored a hypothetical parameter, a "Personal Information Factor" (PIF), which was a measure of the personal information contained in a linked, de-identified dataset or in the outputs of analysis of data.

A PIF above a certain threshold (for example, 1.0) means sufficient personal information exists to identify an individual: the total available personal information makes this personal identifiable. A value of 0 means there is no personal information whatsoever. It is important to note that the PIF envisaged is not a technique for anonymisation; rather, it is a heuristic measure of potential risk of re-identification and the amount of information that would be revealed by re-identification.

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

- A measure of the information content of the dataset used to conduct analysis or the output of the analysis.
- The size of the smallest unique group (which may be an individual) in the dataset or output.
- Additional information required by the observer to be able to identify an individual from the data or outputs (an "epsilon").

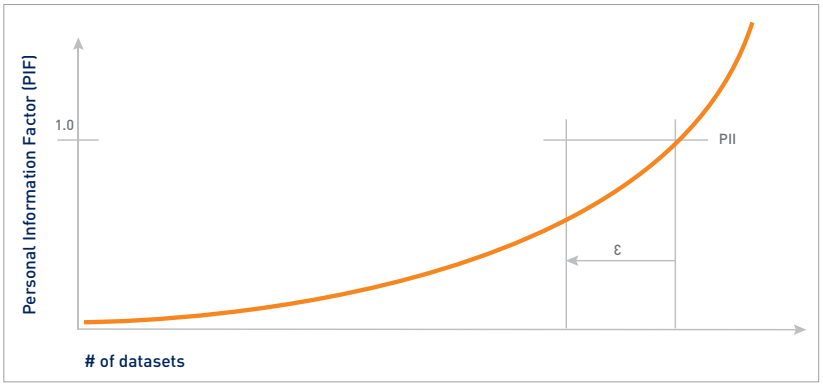


FIGURE 10. CONCEPTUALISATION OF A PERSONAL INFORMATION FACTOR (PIF) AND THE THRESHOLD POINT OF REACHING PERSONALLY IDENTIFIABLE INFORMATION (PII)

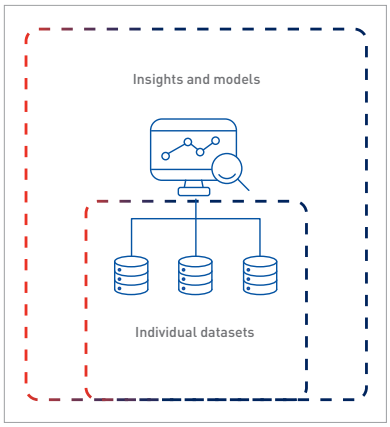


FIGURE 9. CLOSED SYSTEM CONTEXT FOR EVALUATING PIF

Figure 9 shows the context for evaluating the degree of personal information as part of assessing the PIF in a closed system. The data available in a closed (sealed) environment is finite and a PIF can be described mathematically based on knowledge of uniqueness of combinations of features describing individuals.

As an example, consider the analysis of a dataset on passengers who arrive at each train station in a certain region, for each hour of each day, for different passenger types (student, pensioner, adult). Using de-identified input datasets, such analysis may deliver the analytical result (output) that on certain days, at one particular regional station, there is a single person who alights between 3:00pm and 4:00pm and they are a "pensioner". The smallest unique group size is 1 (an individual).

Processing functions may contain embedded extrinsic knowledge such as known probabilities of occurrence of certain values or features. This could increase the PIF of outputs beyond the PIF of the dataset analysed. For example, the extrinsic information added by the analysis is that a "pensioner" must be 60 years or older.

The degree of personal information contained in data may be very high (a unique identifier such as a social security number), moderate (such as surname), low (such as eye colour), very low (such as month of birth) or extremely low (such as weather data).

It is expected that the level of personal information (the PIF) in a linked dataset will generally increase as more, people-centred datasets are linked. As conceptually shown in Figure 10, as more datasets containing PI are linked, a point may be reached where an individual is personally identifiable (a PIF of 1), or "reasonably" identifiable (a PIF within "epsilon" of 1). The dataset is then considered to have PII. The "epsilon" in this figure is an indication of the gap before the "reasonable" threshold is met.

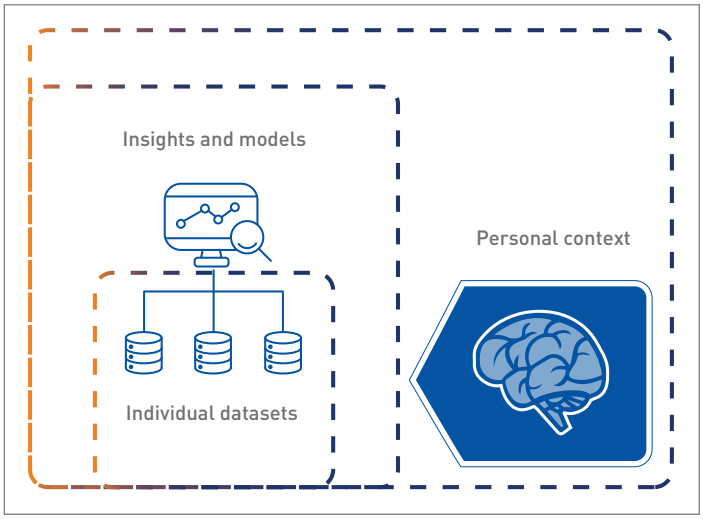


FIGURE 11. HUMAN CONTEXT FOR EVALUATING PIF

Figure 11 shows the context for evaluating the degree of personal information when outputs are observed by individuals who have restrictions placed on their access to data and outputs. The individuals could be screened based on skills, motivation or experience, or may be bound by confidentiality or other legal restrictions. Each observer, however, has their own knowledge of the world or may somehow have a connection to the people represented in the datasets.

Extending the train passenger example above, if the observer has personal knowledge of the regional station identified and knows several people who live nearby and who travel by train, then the PIF associated with insight produced by this analysis is increased for this observer. What the observer does with this increased PIF will depend on the protocols or restrictions placed on them by the environment in which they operate in. For example, they may be removed from the project or may be required to maintain confidentiality of the output.

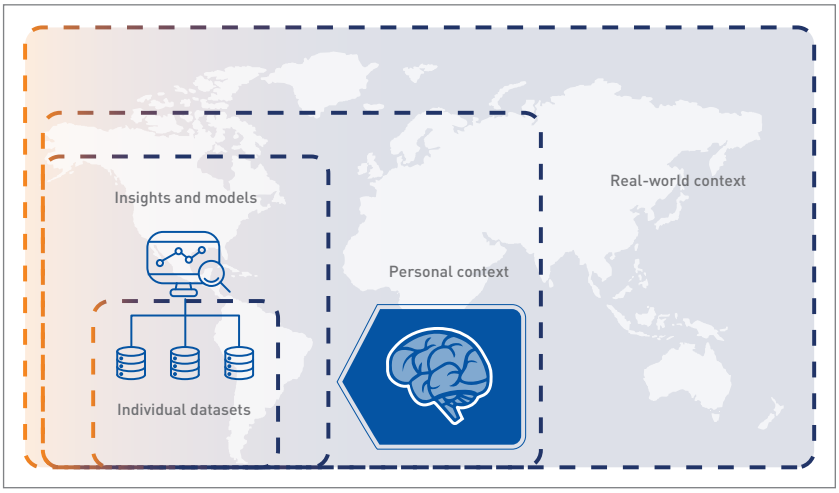


FIGURE 12. REAL-WORLD CONTEXT FOR EVALUATING PIF

Figure 12 shows how a PIF may be considered when project data or outputs have been released into the broadest possible environment, the outside world. Here there is no control over who accesses the dataset or the analysis outputs, and which additional datasets can be combined with the outputs.

Extending the train passenger example above (the observer has personal knowledge of the regional station identified and knows several people who live nearby and travel by train), if this observer waits at the station on the days and at the time the individual is known to travel, then the PIF associated with the outputs can be increased to the point where the individual passenger can be identified. Specifically, the PIF can be brought to 1 (personally identifiable). This requires an observer to expend effort or resources to gain additional information beyond what is made available from the outputs generated by the analysis. If the observer knows that extrinsic information associated with the pensioner rule, they will have the additional information that the identified passenger is over 60.

Personal Information Factor versus risk of re-identification

The premise of the PIF is that, by linking sufficient fields with personal information, a threshold will be reached by which a person can be uniquely (or reasonably) identified (see Figure 10).

Starting with the closed system context for evaluating PIF (Figure 9), and representing individuals as rows in a dataset, with columns as features, the ability to identify an individual means that there are sufficient differences in combinations of column values such that there is at least one unique row *and* there is sufficient information in the columns of that unique row to unambiguously (or reasonably) identify that individual.

In this context, if there are two or more rows with the all the same column values, then there is not sufficient information in the columns to uniquely identify an individual. In this case, the minimum identifiable cohort size (MICS) is greater than one. If there is a unique row, but insufficient information contained in the columns, then the individual can still not be unambiguously (reasonably) identified.

Moving to the human context for evaluating PIF (see Figure 11), the knowledge or experience brought by the observer could increase the total information about a unique row to the point where an individual could be reasonably identified even if there is insufficient information in the columns of the dataset. The observer would need to have knowledge of the members of the dataset (at an individual level or a population level) for this to occur. This could include knowledge that an individual is in the dataset. If two or more rows had the same values, then the observer would still not be able to distinguish exactly which individual was referred to by each row.

Moving to the real-world context for evaluating PIF (see Figure 12), with no control over how many other datasets or other sources of information can be applied to data or outputs of analysis, then there is no absolute protection against re-identification. Relative protections may be applied by limiting the amount of personal information released, and so reducing the information gained by someone seeking to identify and individual, or increasing the effort required to identify by making the smallest number of rows with the same values (the minimum identifiable cohort size) to be a relatively large number.

04

Describing Safe Projects

This paper assumes all analysis is performed using de-identified data. It is also assumed that the de-identified data is not subject to any national security classification.

The question of what makes a "Safe Project" (Figure 13) is one laden with subjective meaning and interwoven concepts. "Safe" is a commonly used term that has a variety of interpretations depending on who is considering the project and what happens to the results. It may be that a project is important but may still deal with sensitive data or address sensitive issues. The results may inform a range of stakeholders and the results used to modify services, develop new interventions or prioritise activities.

The question of a Safe Project is often conflated with issues of:

- **Privacy –** what is the level of personal information present in the data required to undertake the project and what is the level of personal information in the results?
- **Sensitivity –** will the project reveal insights that cause harm or embarrassment?
- **Importance –** are the goals of the project sufficiently worthy to set aside other concerns?
- **Ethics –** what are the consequences of gaining the insights from the project?
- **Outcomes –** what will be the consequences of the insights produced from the project?

People who are involved in deeming a project to be safe are, by necessity, applying a subjective framework to the cost/benefit of a project and are unlikely to have understanding of the long-term consequences.

A "project" in this context is also often conflated with a program of work (supporting an outcome), leading to a lack of clarity as to whether it's an individual analytical task or the larger program it informs that is being assessed as Safe. The Safe level of a project can be considered based on purpose (why the project has been approved to be performed), and what the consequences of the creation and use of the outputs will lead to, including potential harms.

The introduction of the additional Safe dimensions, including Outcomes, Use, and Response in Chaper 3, are intended to remove this ambiguity and allow the consideration of Safe to rest solely with the analytical task to be performed.

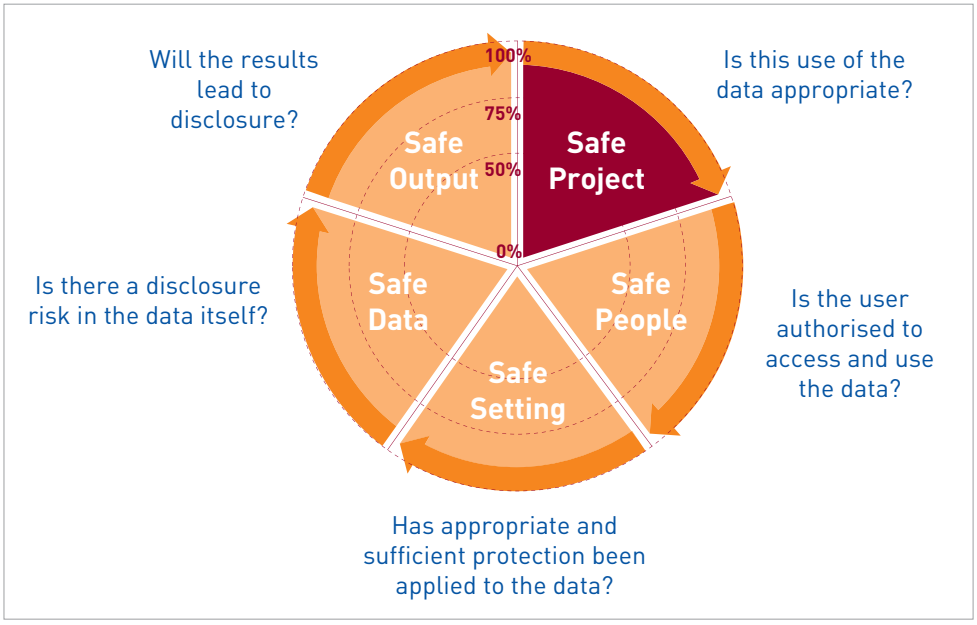


FIGURE 13.
SAFE PROJECT

PRIVACY CONSIDERATIONS

This paper is primarily concerned with the issue of privacy and so, while acknowledging the challenges associated with the use of a culturally loaded term such as "Safe", will seek to explore a Safe Project purely from the perspective of privacy.

From a privacy perspective, a Safe Project can be characterised by the level of personal information in the data required to undertake the project, and what level of personal information is contained in the outputs.

A REAL-WORLD EXAMPLE OF A SAFE PROJECT EVALUATION

Let's return to the local council lake water temperature measurement project.

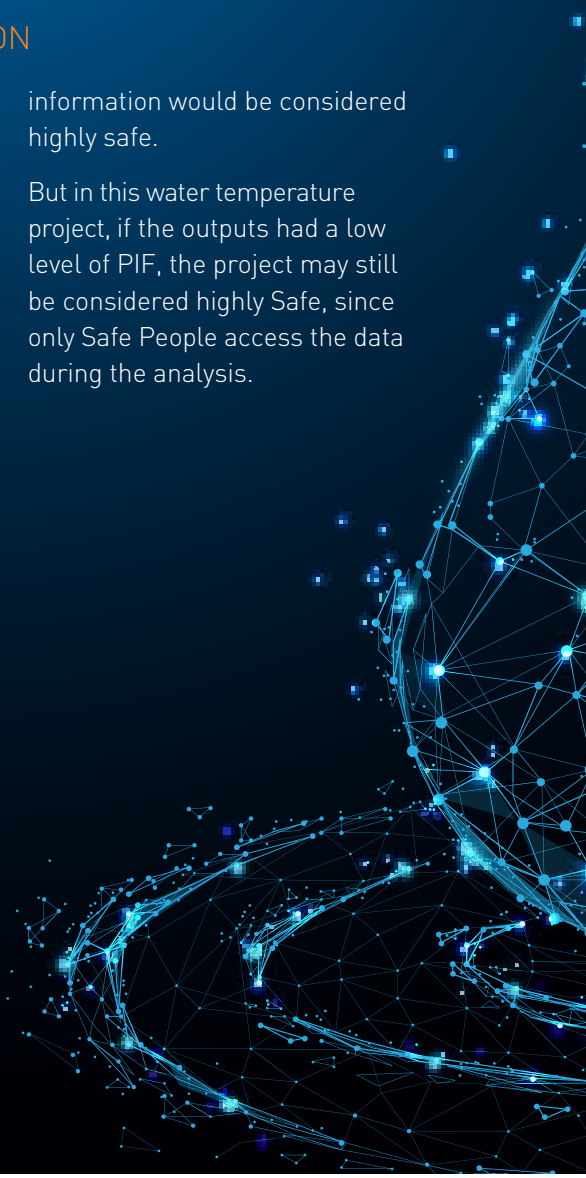
The purpose of the project is to better understand the environmental health of an important body of water. The data are to be collected by temperature sensors, some of which will be located next to isolated dwellings on the lake shore. The data from these sensors are very likely to reveal activity of the people occupying the isolated dwellings. It is therefore considered that this data is likely to have a high PIF, even if de-identified.

The importance of understanding the environmental health of the lake through this temperature analysis is deemed to be a sufficiently important reason to undertake the project. The need to use high PIF data must be mitigated by who access the data ("Safe People"), the level of PIF in the results, and who gets to see the results. In this case, the project should be considered to be not-Safe from a data use perspective, and not-Safe depending on the PIF of the outcomes.

The discussion of the level of Safe for the water temperature project does not address the issues of sensitivity, ethics or outcomes. The motivation for the project is its importance and the assessment of its level of Safe is based on personal information in the data and outputs.

With this perspective, a project that has significant sensitivities, such as a study of domestic violence, but which uses data with low levels of personal information would be considered highly safe.

But in this water temperature project, if the outputs had a low level of PIF, the project may still be considered highly Safe, since only Safe People access the data during the analysis.



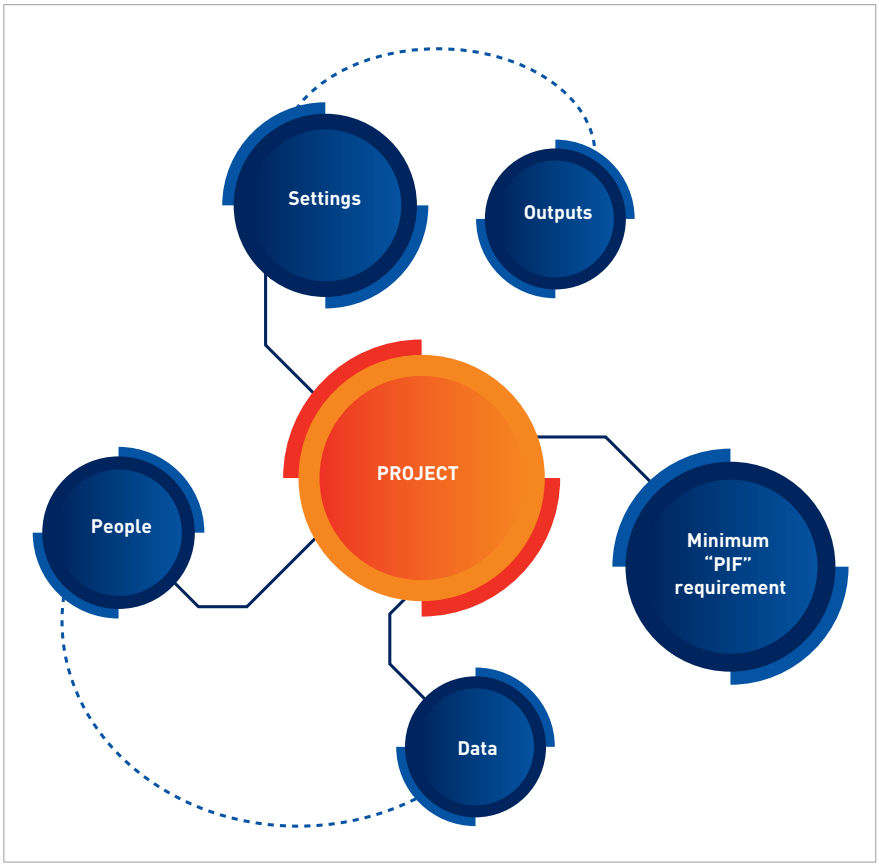


FIGURE 14. INTERCONNECTIONS BETWEEN THE FIVE SAFES.

Exploring the Interactions between risk dimensions shows (Figure 14) that "Project" can drive requirements for "Data", "Settings" and "People". Similarly, "People" can drive "Data" or "Settings", and "Data" similarly interacts with "People" and "Setting". The challenge is to determine which of the dimensions are fixed and which must be adapted for the level of Safe required for the project.

Ethical considerations

By policy or regulation, in many countries ethics approval must be sought for research involving human participants. Formally convened ethics committees exist in most countries to evaluate research projects and to provide guidelines for conduct when carrying out projects. As an example, the UK's Social and Economic Research Council (SERC) provides a framework for research ethics principles, procedures and minimum requirements[7]. These minimum requirements include:

- Research should be designed, reviewed and undertaken to ensure integrity, quality and transparency.
- Research staff and participants must normally be informed fully about the purpose, methods and intended possible uses of the research, what their participation in the research entails and what risks, if any, are involved.
- The confidentiality of information supplied by research participants and the anonymity of respondents must be respected.
- Research participants must take part voluntarily, free from any coercion.
- Harm to research participants must be avoided in all instances.
- The independence of research must be clear, and any conflicts of interest or partiality must be explicit.

There are additional requirements for vulnerable groups or sensitive topics. The SERC guidelines are typical of many ethical frameworks and require judgement from an expert panel in the event that issues are identified with the project.

When evaluating Safe Projects from an ethical perspective, two threshold questions arise:

- Does the de-identified people-centric dataset qualify as research (analysis) on humans?
- Can the full scale of potential issues be identified upfront, before the project commences?

The first point is dependent on understanding the degree of personal information the de-identified dataset. There are many views within the research community as to whether deidentification drops a potential project below the ethics-seeking requirement. The 2018 ANDS guide[8] on data sharing for human research states:

Under the Privacy Act 1988, sensitive human and personal data cannot generally be shared in their original form. However, once de-identified, these modified data no longer trigger the Act as they are not 'personal information'. In other words, de-identified sensitive data can legally be shared.

The guide further states:

It is worth noting that whilst the Privacy Act 1988 does not apply to de-identified data, it does apply to the activity of de-identifying the data (i.e. removing identifying information from the original, sensitive dataset), and it might also apply in the context of seeking to re-identify data. This activity is, however, explicitly condoned in the Australian Privacy Principles of the Privacy Act 1988 as one of few exceptions to sensitive data use. This is because de-identification is considered a 'normal practice' that 'an individual may reasonably expect their personal information to be used or disclosed for' without requiring specific consent.

The second issue led to the suggestion in the 2017 white paper to develop evolutionary governance processes, wherein the level of governance applied to People, Setting or Output could be evolved as the project progresses. Once again though, the issue is primarily one of having crossed the threshold of personal information factor which would allow for re-identification.

[7] Available online http://www.gla.ac.uk/media/media_326706_en.pdf

[8] See *Australian National Data Service Data sharing considerations for Human Research Ethics Committees*, June 2018, Available online https://www.ands.org.au/__data/assets/pdf_file/0009/748737/HREC_Guide.pdf

Sensitivity considerations

The discussion of sensitivity in Chapter 3 highlighted a number of factors that describe the sensitivity of a project and thereby the risk factors that need to be addressed. The sensitivity of a project is driven by the purpose of the project (outcome), the data planned to be used, the scope of release of the outputs, how much context is required to interpret the outputs, and how the outputs will be used. Beyond privacy concerns, the sensitivity considerations will also drive selection criteria for people who work on the project, and the setting the project is carried out in.

Figure 15 shows an example governance framework for projects with different levels of sensitivity and privacy. In the bottom left hand corner, a project with Low Sensitivity and using data with low levels of personal information may be undertaken with very little pre-approval and the results released widely – a project carried out by a member of the public using open data, for example.

On the bottom right-hand side, a project relying on data with high levels of personal information but Low Sensitivity may be carried out with appropriate controls over personal information and outputs may be released widely if appropriate protections to manage re-identification risk. Such projects are carried out regularly by organisations such as the Australian Bureau of Statistics with release of aggregated Census information.

In the top left-hand corner, an analysis which is highly sensitive but relies on data with a low personal information will require expert review and may have some restrictions on the use of outputs. In the top right-hand corner, high levels of governance and expert review are required at all stages of the project.

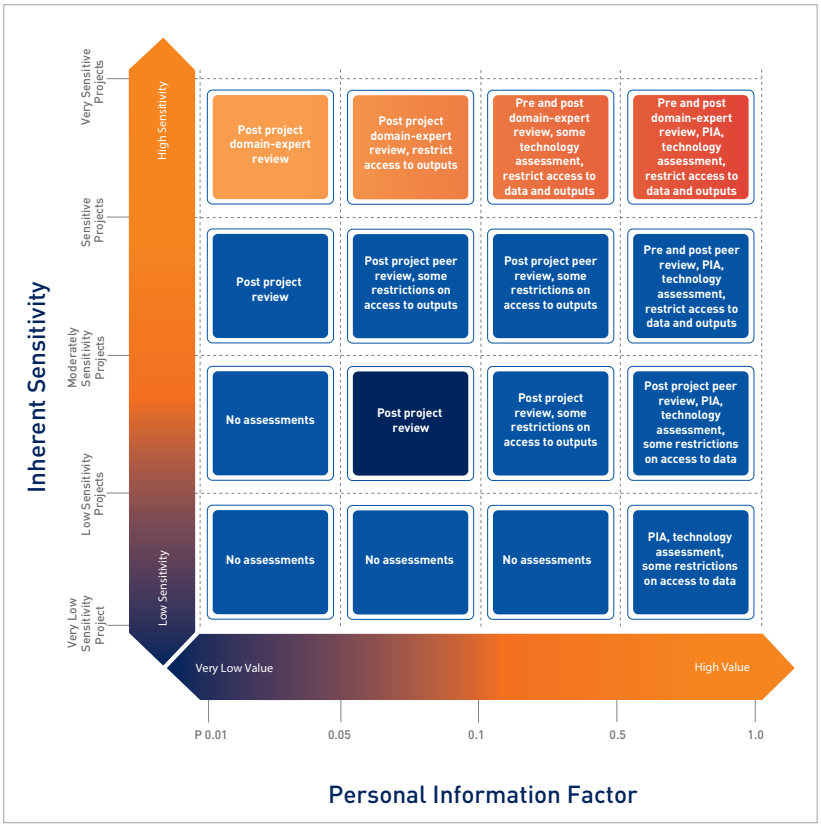


FIGURE 15. EXAMPLE GOVERNANCE FRAMEWORKS FOR PROJECTS WITH DIFFERENT LEVELS OF SENSITIVITY AND PRIVACY

05

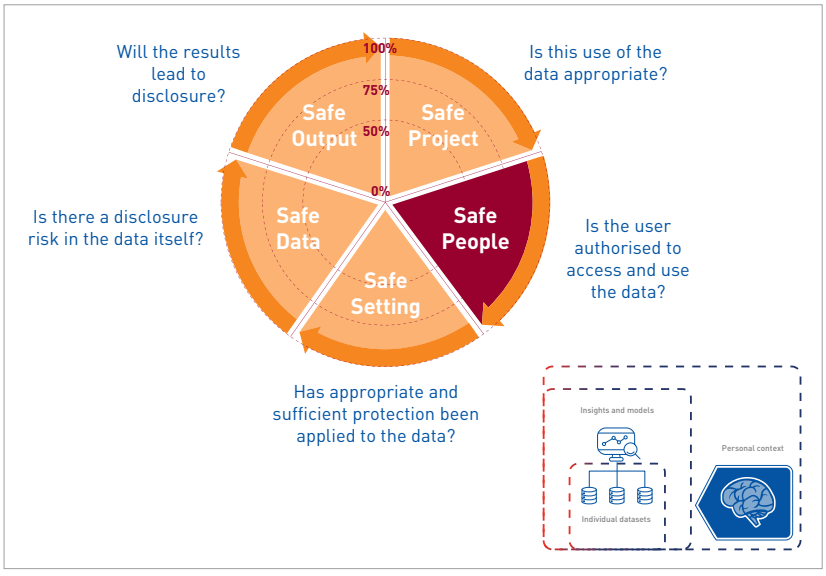
Describing Safe People

Like Safe Projects, the concept of Safe People (Figure 16) is a similarly overloaded concept. In a research or analytics context, a Safe Person may be considered to be someone who is:

- Skilled in analytical techniques.
- Screened or endorsed by independent authorities.
- Bound by legal agreements or formal undertakings.
- Appropriately motivated to perform the project.
- Not connected to individuals represented in the dataset.

These attributes seem intuitively reasonably and may be used to describe a "Very Safe" person who could potentially work on very unsafe projects. The challenge, however, is being more specific about what is meant by Safe Person.

FIGURE 16. SAFE PEOPLE



There is the second issue that, despite all undertakings and training, the motivation for a person to undertake a project can only truly be known by the person undertaking the project. Employment or financial reward is one motivation; however, for many people, there are other motivations. In extreme cases, such as the WikiLeaks[9], data and outputs have been deliberately leaked motivated by a sense of social justice. Someone who works for a commercial analytical firm may also be motivated by personal reward to push closer to re-identification of individuals in order to, for example, create more effective personalised offerings.

Someone who is skilled in analytical techniques may be more likely to produce a quality output than someone who is not. From a privacy perspective, however, someone more skilled in analytical techniques is more likely to be able to re-identify an individual than someone who is not skilled. From this perspective, they could be considered less Safe.

Similarly, someone who has a relationship with the individuals represented in the dataset is more likely to be able to identify someone in the dataset. This makes them less Safe from a privacy perspective.

[9] See, for example, https://en.wikipedia.org/wiki/WikiLeaks

Privacy considerations

Viewing the dimension of Safe Person from a privacy perspective ultimately centres on the ability to re-identify individuals from de-identified datasets, the systems to minimise that risk and the consequences in the event that it does occur.

An alternative definition for a Safe Person may involve measuring:

- **Skill in data governance –** rather than analytical capability, accredited skill in being able to take appropriate steps when handling data at different stages of the project lifecycle.
- **Personal connection to the dataset –** understanding the degree of separation between the people represented in the dataset, or the region represented in the analyst.
- **Accountability –** the consequences for the analyst in the event that re-identification does occur (PII is attained), PII is released, or PII is used inappropriately by the analyst.
- **Organisational capability –** an analyst operating in an organisation.

The consideration of personal connection to a dataset reflects the example, provided in Chapter 3, of the pensioner alighting at the regional station. The more independent knowledge an analyst has of the members in a particular dataset, the more opportunity they have to re-identify individuals.

Measurement of "connection" to the members of a dataset needs to be done with care. Connection could be by personal features (people with similar health conditions), personal relationship (friend, family, classmate, colleague), spatial relationship (neighbour, suburb, town, state) or even temporal (birthdates, hospital admission dates).

Being able to make direct comparisons between an analyst with personally identifiable information of the individuals in the dataset would allow ready assessment of personal connection. It would, however, be counterproductive when considering privacy and the focus on de-identified datasets. Being able to systematically generalise datasets by categories of personal features, relationship features, spatial or temporal features would allow an analyst with concerns about association to be able to use more generalised data. This, however, would be a "Safe Data" response to a less Safe analyst.

Sensitivity considerations

A Safe Person may be described as someone who recognises the appropriate governance levels required as a project develops and will apply the appropriate governance accordingly irrespective of the sensitivity of the outputs at the different stages of the project lifecycle.

Similar to Figure 15, Figure 17 shows an example of a governance framework for people working on projects with different levels of sensitivity and privacy. Approaching the top right-hand corner, there are increasingly advanced reference checks, assessment of analytics and governance capability and assessments of connection to the datasets being analysed. In this diagram, the different dimensions of sensitivity are not broken down.

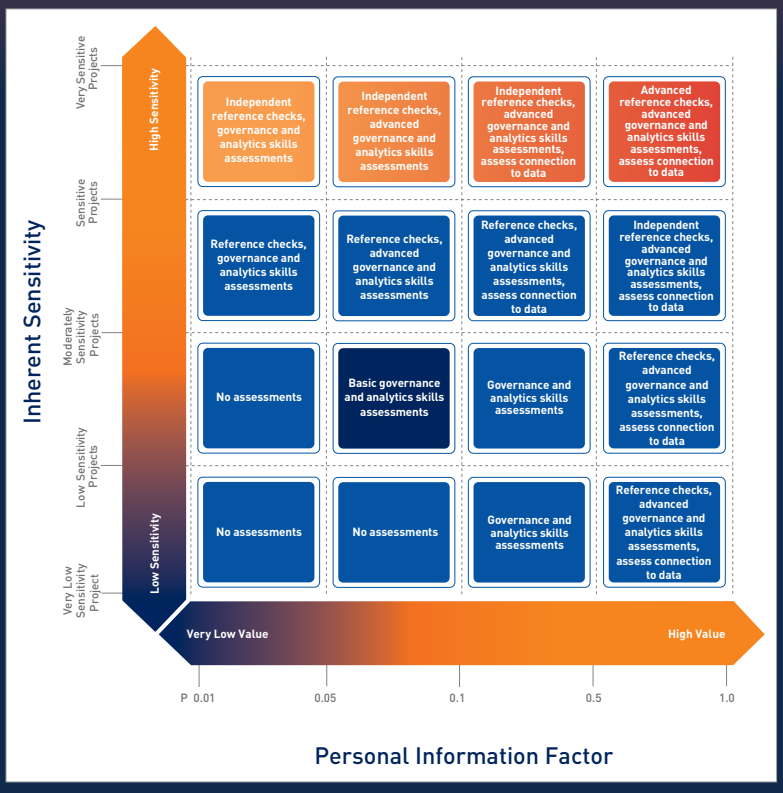


FIGURE 17. EXAMPLE GOVERNANCE FRAMEWORKS FOR PEOPLE WORKING ON PROJECTS WITH DIFFERENT LEVELS OF SENSITIVITY AND PRIVACY

06

Describing Safe Data

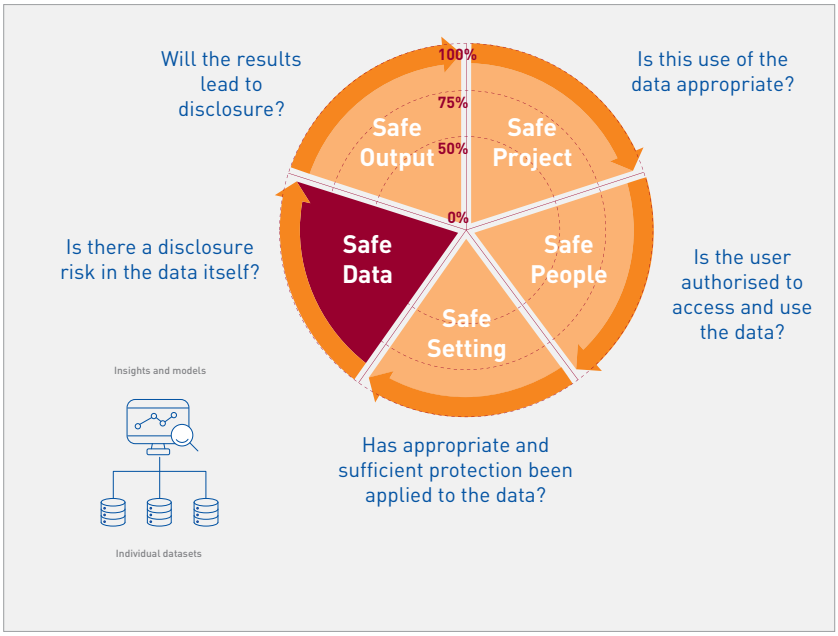
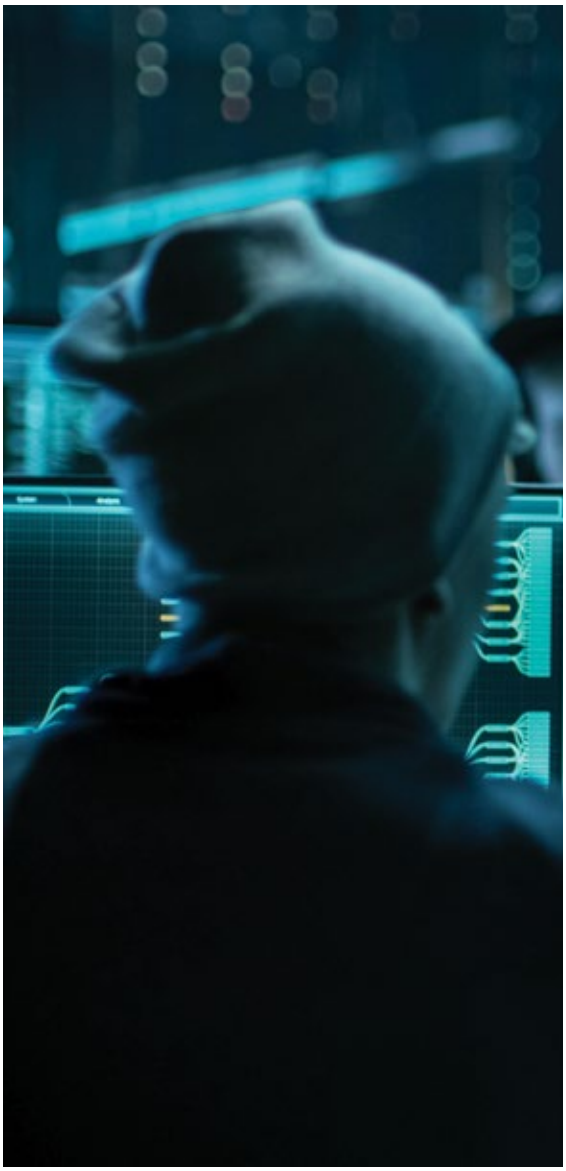


FIGURE 18. SAFE DATA ENVIRONMENT

The challenge of determining a personal information factor (PIF) is ultimately the challenge of understanding how data is to be used, who will have access to it, and what other datasets would be required to re-identify an individual. This potentially reintroduces all of the dimensions of the Five Safes framework. One approach is to focus initially on the dataset itself and its properties, decoupling this risk dimension from all other dimensions, as shown in Figure 18.

Determining a meaningful PIF for a closed analytical environment as described in Figure 9 is the most straightforward scenario, as there are strict controls on which datasets are analysed and on outputs. For simplicity, it is initially assumed that no extrinsic information is added by processing.

During February 2019, ACS ran a Directed Ideation event where teams competed to progress ideas to develop a PIF, build risk frameworks for data based on these PIF models and to attempt to generate datasets of different Safe levels from three base datasets. This approach was repeated in July 2019 to extend the measures of PIF and explored the utility of datasets that have been protected through aggregation. The datasets used in the July event are described in Appendix A and are referred to throughout this document as examples.



An example approach — information gain

A promising approach is to base the PIF on a quantified measure of the information gained by re-identification of an individual. The approach is based on concepts of information theory and cryptography. The approach is summarised in Figure 19. This is not meant to be a definitive approach to defining Safe Data, but has shown promise in explorations to date.

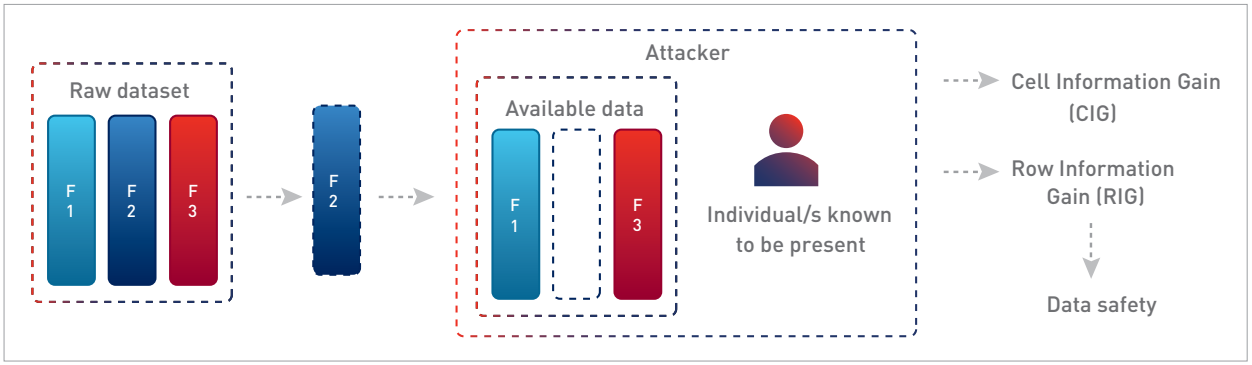
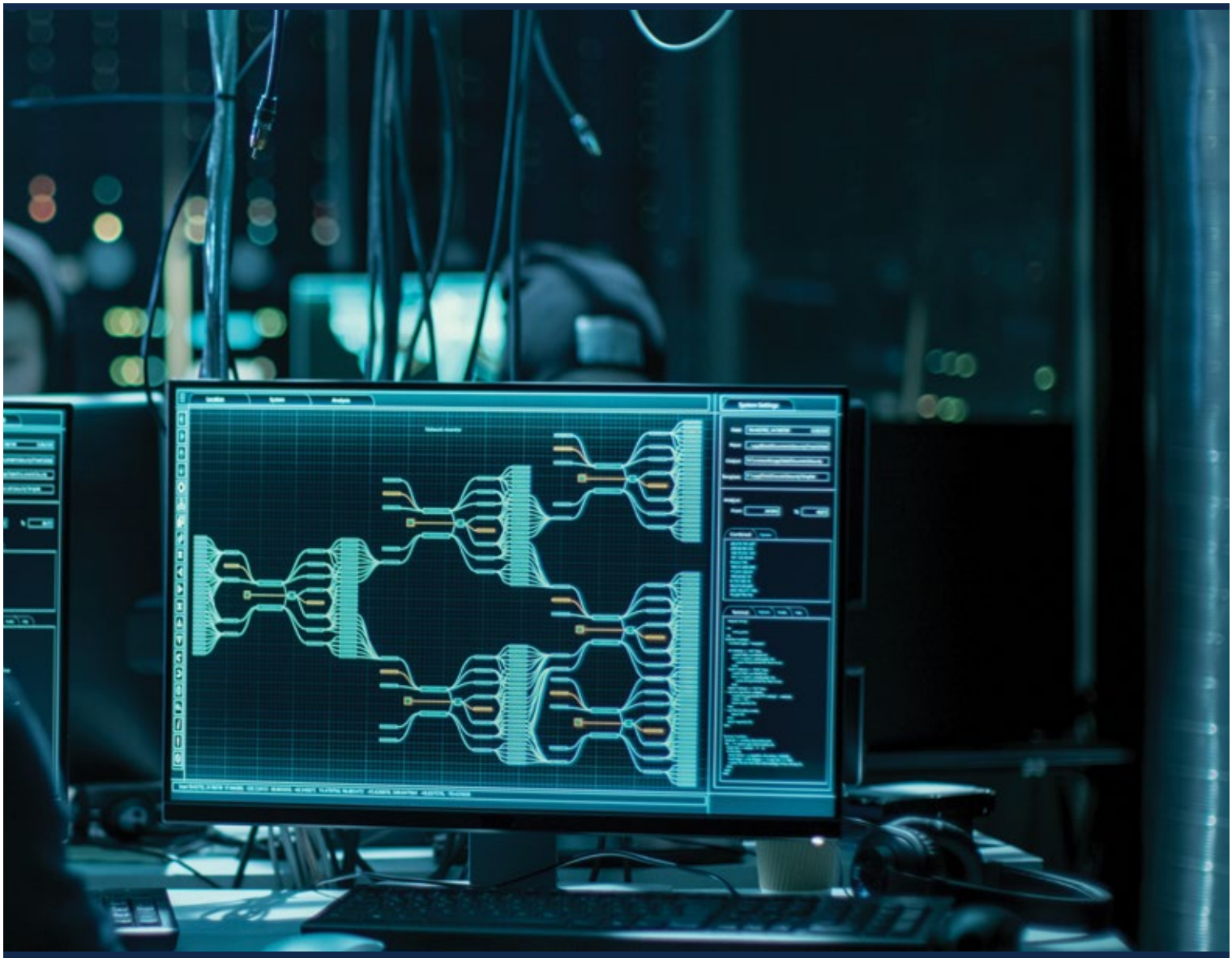


FIGURE 19. OVERVIEW APPROACH



The approach is motivated by the fact that every dataset released into a wider world is available to an "attacker" seeking to re-identify someone. Every dataset therefore represents an increase (possibly small) in the information available to the attacker about a person whose details are described in the dataset.

It is important to note that not every re-identification event is equal in terms of the personal information revealed. For example, learning that a person's 2018 income was in the range from $50k to $150k reveals less personal information than learning the exact 2018 income. So instead of focusing on the individual's re-identification risk, the PIF framework computes the potential information gain for each field in the dataset.

For simplicity, the approach assumes that each row of a dataset represents an individual, and each column represents a feature.

The framework then allows the data steward to:

- Reason about risks on a per-feature (column) basis.
- Find individual risk of each person (row).
- Identify comparatively high-risk individuals.
- Prioritise anonymisation efforts to focus on the most vulnerable features and individuals.
- Compare the performance of different anonymisation strategies.

THREAT MODEL

The approach uses a model from cryptography to formalise the threat from an attacker. An attacker is a person who has access to the dataset and to additional information about an individual they are seeking to re-identify. By locating and re-identifying an individual in the dataset, the attacker seeks to learn more about them.

Knowing the information and resources (the strength) of an attacker is difficult, as the auxiliary information available to the attacker is unknown. Consequently, the approach assumes different strength of attackers when access to data and results are controlled by technology and process, and a very strong attacker when there is no restriction on access to data or processing resources.

A very strong attacker is defined as knowing every feature of a person aside from the one they are attempting to find. Less strong attackers are described as those who know some but not all features, or those who are not fully certain in the information that they have.

Quantifying the amount of information the attacker learns about a re-identified individual remains to be defined.

QUANTIFYING INFORMATION CONTENT

The quantification of information is well known in information theory and described in units of 'bits'.

An information theoretic definition states that the amount of information associated with a given value being generated by a random process is inversely related to the probability of that value occurring. As an alternative description, the less likely a particular value is of occurring, the more information associated with the occurrence of that value. The number of information 'bits' is then the logarithm (base 2) of the inverse of this probability.[10][11]

For example, a coin toss of a fair coin is a binary choice where both options are equally likely, so each coin toss 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.

This approach to quantifying information has been used for more than 70 years to analyse the information associated with communications systems. In 1948, Claude Shannon published his landmark paper, "A Mathematical Theory of Communication" in the *Bell System Technical Journal* Shannon showed how all recorded information could be quantified with precision and demonstrated that information media – ranging from telephone signals, text, radio waves or pictures – could be encoded as digital bits and transmitted at a known maximum rate over a channel.

The information theoretic model has been applied in ever-expanding fields of information media that can be represented in data. The theoretical frameworks developed are however only strictly applicable when a data source is well defined, and the communication channel can be accurately characterised.

We can now build on this concept by using (Kullback–Leibler) divergence calculation to produce a measure referred to as the Cell Information Gain (CIG), a Row Information Gain (RIG) and a Feature Information Gain (FIG).

[10] See, for example, R. M. Gray, *Entropy and Information Theory*, Springer-Verlag, 2014. Available online http://ee.stanford.edu/~gray/it.pdf

[11] Importantly, another fundamental of information theory states that additional processing of data will not create additional information beyond what is already present. This is significant when considering the limits of analytical models.

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 assume a 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 Figure 20.

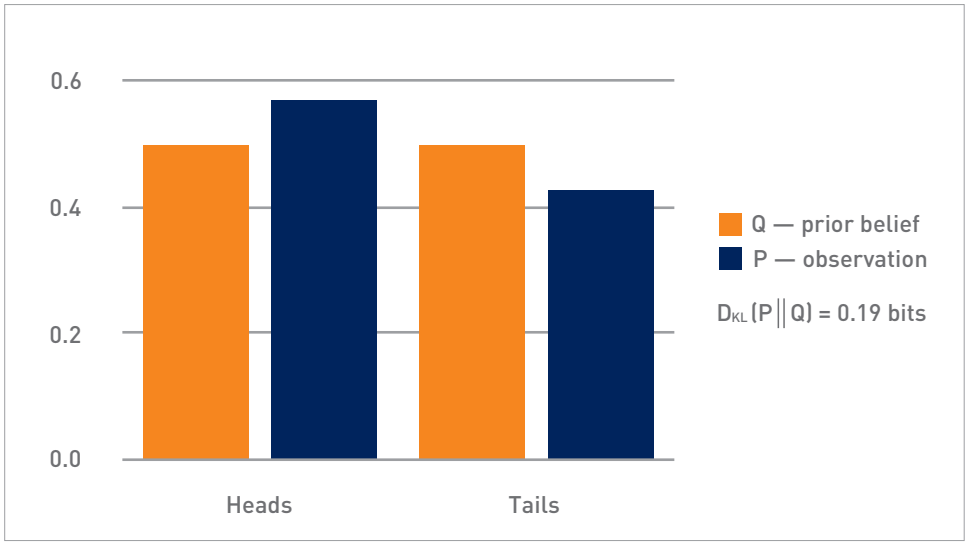
Conversely, if upon investigation we find that our coin is unbiased, the KL divergence of our prior and posterior probability distributions is 0 bits. This is because the probability, as estimated by us, of the outcome was unchanged.

When discussing re-identification risk, 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 all possible birthdays equal probability (excluding February 29).

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.[12] If the attacker narrows the birthdate down to two equally likely options, then the information gain is 7.5 bits.[13] If the attacker learns nothing, then the KL divergence of the prior and the posterior is zero.

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

FIGURE 20.
KL DIVERGENCE OF THE COIN TOSS EXAMPLE



[12] The change in probability of birthday goes from 1 in 365 (all equally likely) to 1. The subsequent information gain is $\log_2(365)$ which, to one decimal place, is 8.5 bits.

[13] The change in probability of birthday goes from 1 in 365 (all equally likely) to 1 in 2. The subsequent information gain is $\log_2(365/2)$ which, to one decimal place, is 7.5 bits.

CELL INFORMATION GAIN

Returning to the attacker model, the Cell Information Gain (CIG) is defined to quantify the re-identification risk for each feature for each individual. Once again, for simplicity, it is assumed that every cell belongs to a row, and every row represents information about a person.

Considering a strong attacker scenario, for each cell we wish to determine the CIG for, we assume that the attacker knows every feature of the person they are seeking to re-identify, other than the cell being considered. The CIG value 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. In the strong attacker case, we almost certainly do not have access to this prior. We can approximate this prior within the dataset by tallying the occurrences of every possible value of this feature across the entire dataset.

In a similar way, we can calculate the posterior, if we assume there is a particular individual the attacker is targeting, and we have a vector (row) of features for this person. For every person (row) in the dataset we assign a probability that they are the person the attacker is seeking to re-identify. For every possible value of our cell, we tally the occurrences of the people (rows) who have this value. This calculated posterior is compared with the prior to give us our CIG in bits.

The calculated CIG (in bits) is given by:

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

CALCULATING INFORMATION GAIN IN PRACTICE

The overall approach to processing data and calculating the various information gains is:

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

High FIG features or high RIG rows may be preferentially targeted for suppression, aggregation or other

FEATURE INFORMATION GAIN

It is now possible to build on the CIG. By summing the values for each feature, we can find the Feature Information Gain (FIG) for that feature in the dataset. The FIG is therefore a measure, in bits, of the re-identification risk of that feature. It can then be used to identify the features that are the highest risk to include in a dataset. 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

In a similar way to how the FIG was determined, we can build on the CIG to develop an information gain for every row, or individual, of the dataset.

The Row Information Gain (RIG) is determined by summing all the CIG values in the row and is a measure of how susceptible a particular individual is to having their information revealed through re-identification.

forms of protection to reduce information gain when data is shared or released.

For example: consider a sample of the Hospital Admissions dataset (see Appendix A, dataset 7) with CIG values shown in Figure 21.

In this dataset, all rows have large information gain for the "POSTCODE" feature, making this a relatively high-risk feature to include in the dataset if shared. Row 6 also has relatively large information gain for the "job" feature, making this individual relatively high-risk to include if the data is shared.

When considering how to de-risk the dataset for sharing (make the dataset "Safer"), large CIG, RIG or FIG values can be altered or removed to reduce the PIF of a dataset.

	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	11.3243	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	10.0612
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	2.54684	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	11.4238	2.99185	2.33236	4.85394
14	1.32193	2.54684	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	9.22647	1.99582	2.33236	3.50535
18	0.736966	3.51905	5.21259	2.97789	2.31098	2.60658
19	0.736966	3.54684	11.9092	3.01561	2.31117	5.157

FIGURE 21. EXAMPLE FICTITIOUS "MEDICAL" DATASET

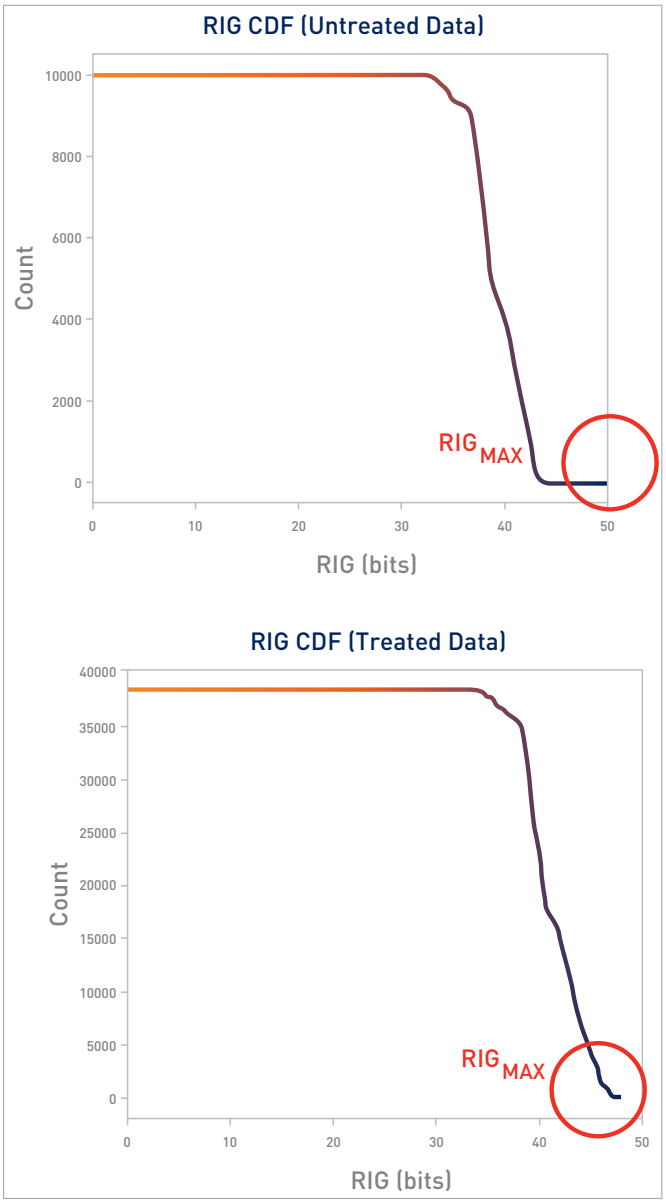


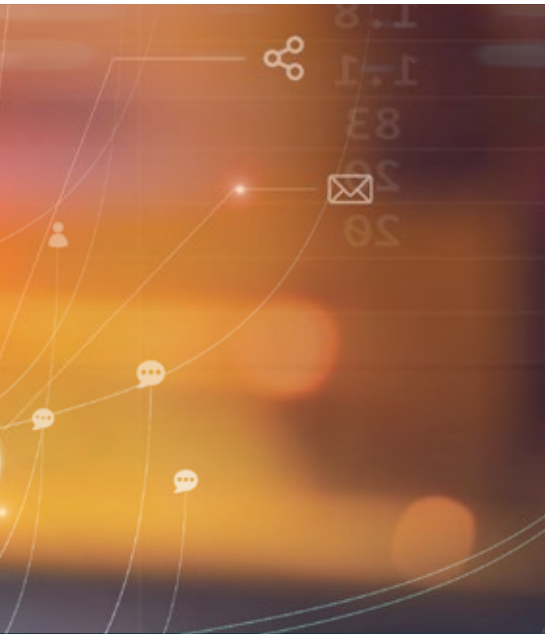
FIGURE 22. RIG SCORE DISTRIBUTION BY DATASET

Linking PIF to re-identification risk within a dataset

The information-gain based approach allows the re-identification risk of the dataset to be expressed in a number of ways.

Figure 22 shows the distributions of individuals' RIG levels in two synthetic datasets. The horizontal axis shows the count of individuals with a RIG level higher than each RIG threshold (in bits). The dataset on the left has a lower average for all RIG values than the set on the right. However, the left-hand dataset has a small number of individuals who are at an elevated RIG level, which is shown by the long tail in the bottom right of the plot.

Despite having a lower average of all RIG values, the dataset on the left has a higher absolute risk of



re-identification due to the small number of high information gain rows. Depending on how the data is treated – deleting, perturbing or aggregating values in cells – a much safer data product can be created from this dataset without significantly impacting the majority of cells and rows.

By defining the quantity RIG_{95} as the 95th percentile of all of RIG values, it is possible to characterise the re-identification risk of the entire dataset in a single number. Similarly, we can define as the RIG_{max} of the individual about whom the greatest amount of information would be revealed if re-identified.

The PIF for the dataset is driven by both the smallest identifiable cohort size (MICS) and the amount of information that would be revealed if individuals in this cohort were re-identified. The definition of PIF is still a work in progress, but the current working definition is given as:

$$PIF = \text{maximum of } (RIG_{(x)} / (MICS \text{ at } RIG_{(x)}))$$

At any given RIG threshold, the MICS at that value is the smallest number of rows with all the same column values. For example, if the number of rows with a RIG at RIG_{max} is 1, then the PIF is equal to RIG_{max}. If the number of rows with a RIG of RIG_{max} is 2, and there are no other unique rows in the dataset, then the PIF is $RIG_{max}/2$. If there is a unique row at a threshold RIG less than RIG_{max} (say $RIG_{(X)}$) and the number of rows at is RIG_{max} is 2, then the PIF is $RIG_{(X)}$, provided $RIG_{(X)}$ is greater than $RIG_{max}/2$.

It is stressed that this is a working model of PIF and is yet to be robustly tested. Example thresholds for different Data Safe levels:

Safe Level 1: 1.00 ≤ PIF

Safe Level 2: 0.33 ≤ PIF < 1.00

Safe Level 3: 0.11 ≤ PIF < 0.33

Safe Level 4: 0.04 ≤ PIF < 0.11

Safe Level 5: PIF < 0.04

Setting a threshold of Safe Level 1 as a PIF greater than or equal to 1.0 acknowledges the fact that the least Safe data is reasonably likely to be able to re-identified. Safe Level 5 should be sufficient for release of data to the outside world as open data. Worked examples of PIF settings with aggerated datasets will be shown in Chapter 12.

Extending the Information Gain Framework

The Information Gain Framework makes it possible to extend and adapt to a wider range of use-cases.

FEATURE ACCURACY

Data accuracy (or individual feature accuracy) is a complex challenge when considering risk of re-identification. Re-identification of an individual based on a set of features, subsequently leading to gaining knowledge of a set of additional features (not used to identify) that are inaccurate minimises the information gained by an attacker, but may lead the attacker to believe information has been gained.

Similarly, incorrectly identifying an individual due to inaccurate feature values may lead the attacker to believe information has been gained about the "wrong" individual. Injecting noise values into a dataset is often used to reduce the risk of re-identification at the cost of reducing the accuracy of the dataset.[14]

The framework for calculating the CIG allows inclusion of the level of noise (or inaccuracy) in the dataset. This affects both of the posterior computations for the cell. The approach takes the uncertainty into account when assigning to each member of the dataset 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, the higher uncertainty in the data the lower the CIGs.

[14] See, for example, the Australian Bureau of Statistics https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%2Subject/2011.0.55.001~2016~Main%20Features~Data%20Quality%20and%20Random%20Perturbation~18

FIGURE 23. IMPROVED CIG MEASURES USING KNOWLEDGE OF POPULATION DISTRIBUTIONS

	gender	AGE	POSTCODE	blood_group	eye_color	job
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	AGE	POSTCODE	blood_group	eye_color	job
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

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. Figure 23 shows the CIG for elements of the hospital admission dataset without (LHS) and with (RHS) knowledge of the distribution of the feature "icd_code".

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

USE OF DIFFERENT ANONYMISATION TYPES

The described technique for calculating CIG is agnostic to the kind of anonymisation used. A common technique for anonymisation is k-anonymity. Another approach may be to perturb the values before release. In this case, we assign an accuracy value 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 ATTACKER CAPABILITIES

The most conservative approach is to assume the attacker is very powerful; that is, they know every feature of the person they are attempting to re-identify except for the one feature they are attempting to find. This most conservative approach is relevant for the real-world context as described in Figure 12.

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, it is possible to assume the attacker knows *n* features of the individual 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 re-identifying them. This is reflected by lower CIG (and consequently FIG and RIG) scores across the dataset.

Another possible attacker model 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 described.

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 re-identification risk calculation.

A major challenge – dealing with trajectories

The discussion of PIF or information gain has a tendency to look at "human" features as being the key to identifiability. This view fails to recognise the impact of spatial, temporal and relationship features increasing the possibility of re-identification by an attacker.

The solution is not always just to aggregate, as time and space have more dimensionality, giving more context that may allow re-identification an individual, or linkage to other datasets. The trajectory approach considers the combined impact of persona, spatial and temporal features.

The "trajectory" of an individual is defined as the set of all rows pertaining to this subject (presumably linked by a dataset identifier or study ID), which describes the longitudinal journey of this individual and their interactions with the dataset.

In the same way as processing operations on data may contain extrinsic information that increases the PIF, it is possible that there may be identifiable properties of the trajectory itself that increases the PIF beyond that found in the dataset itself. For simplicity, it is assumed that this is not the case.

In a longitudinal dataset, each trajectory may be unique (many variables in space-time), and highly different so each individual is potentially identifiable. Consideration of similarity of time-space trajectories as well as personal features can potentially provide a more robust measure of re-identification risk.

Some of the assumptions of this approach include:

1. No matter how mundane a given row is within the dataset, the presence of (and relationships between) multiple rows for a given subject is highly likely to be unique across time, space or both. We therefore define a trajectory that considers all rows for a given subject, across all requested measures of space-time.

2. There exists some transform '*T*' that can project all trajectories in a dataset into a space where a meaningful distance metric can be applied. This allows the creation of a measure of heterogeneity of trajectories across the dataset. We propose that the more heterogeneous a dataset is in terms of its trajectories, the higher likelihood of a possible match when linking to a secondary dataset.

3. In a particular context it is possible that the relative time-space changes may matter as much as the absolute time-space values. This means that it may be insufficient to calculate the uniqueness of a set of values without considering the derivatives of these values (for example, location -→ location change -→ location change velocity), depending on the anonymisation techniques applied.

A REAL-WORLD EXAMPLE OF TRAJECTORIES

Netflix famously released a large de-identified dataset for use in the Netflix Prize.[15] This dataset was soon re-identified by cross-referencing users who rated similar sets of movies in IMDB. The re-identification confidence increases as the number of movies rated on IMDB increased (higher chance of a unique match) – therefore it is not the uniqueness of the rows in and of themselves providing the linkage confidence, rather the uniqueness of combinations of rows, or trajectory. This confidence increases again if it is possible to link behaviours within a given time period, as the combination of factors gains more dimensions and the statistical likelihood of a match can become more precise.

[15] See, for example, https://www.thrillist.com/entertainment/nation/the-netflix-prize

Evolving a PIF to include the impact of trajectory based on the contention that while a static table (see Figure 24) will highlight highest-risk rows, it is the uniqueness of the trajectory from linked rows in a longitudinal dataset that makes an individual at higher risk of re-identification (see Figure 25).

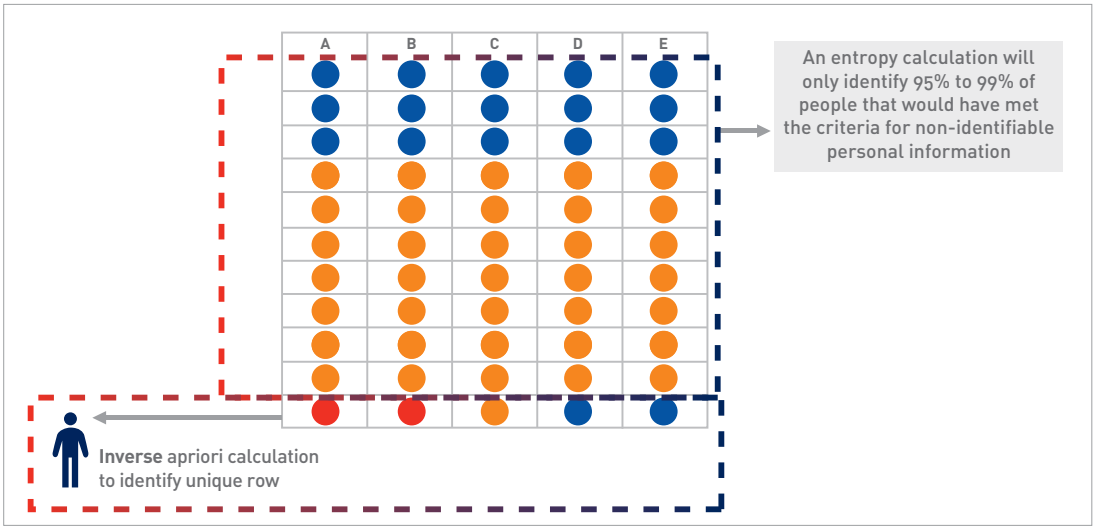


FIGURE 24. UNIQUENESS IN A STATIC TABLE

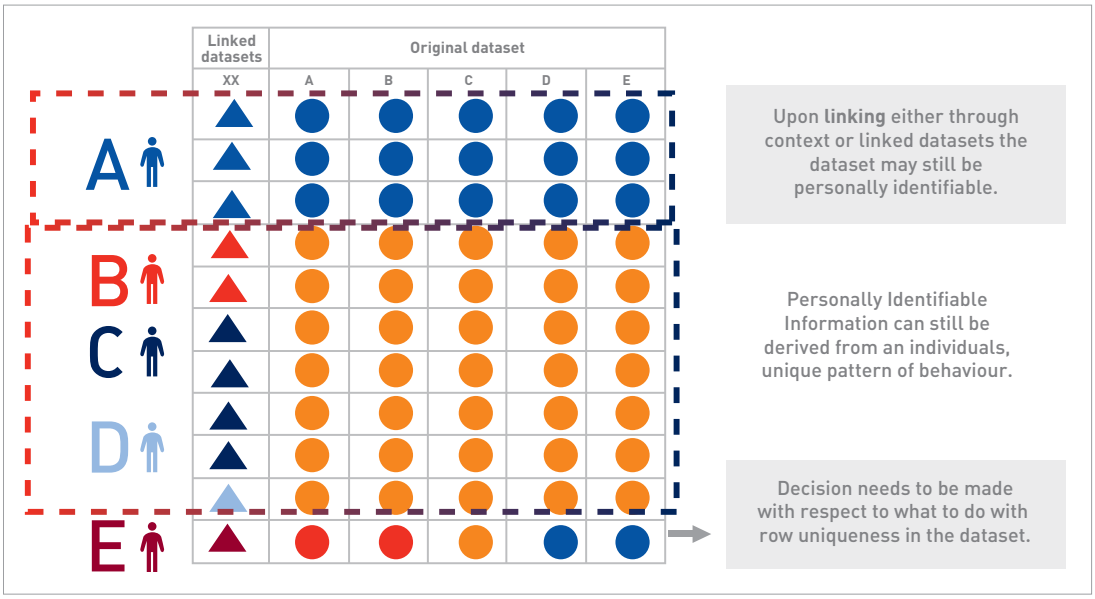


FIGURE 25. UNIQUENESS BASED ON TRAJECTORY

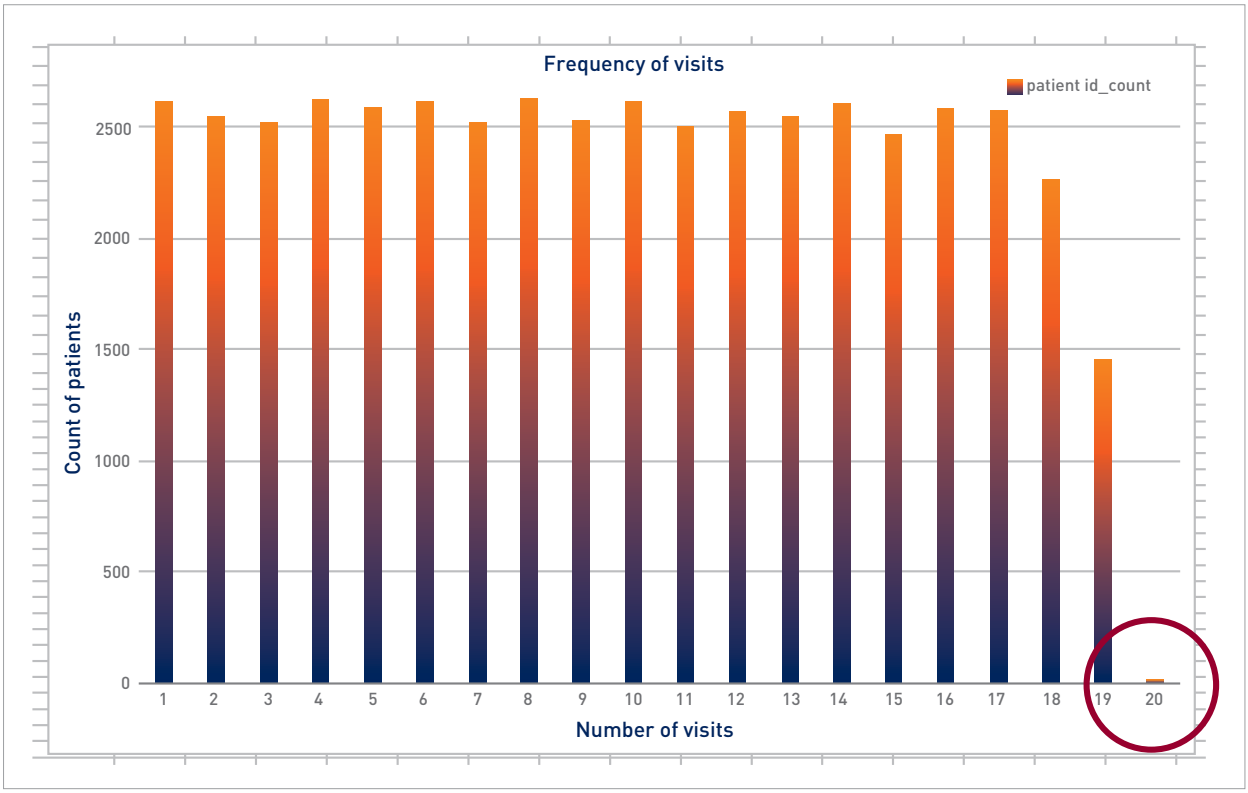


FIGURE 26. UNIQUE TRAJECTORY IN DATASET 7 (HOSPITAL ADMISSIONS)

For example, in the sample dataset 7 (Hospital Admissions), there was one clearly unique trajectory: one individual with 20 admissions (see Figure 26).

The approach is to define a measurable trajectory concept that has a meaningful distance metric, specific to an individual's behavioural patterns.

THE CALCULATION OF A DISTANCE METRIC SHOULD FOLLOW THESE STEPS:

STEP 1

Define a rule-based set of engineered features used to define a trajectory based on time, space or relationship features.

STEP 2

Define a distance measure that describes the heterogeneity of trajectories for a given dataset. This measure should be able to measure the heterogeneity of the whole dataset, and also the distance of any outliers (defining outliers either by their relationship to the whole set or their relationship to some detected clusters).

STEP 3

For each subject, define a t*n matrix where t is the number of rows for this subject and n is the number of space-time-relationship measures for a given dataset (including engineered features such as combinations of features).

Time, space, personal features and relationship features

When thinking of how to protect data, aggregation, suppression or perturbation can be applied equally to the entire dataset, or preferentially to temporal, spatial, personal or relationship features. The intention is to maintain utility in one of more of these feature domains while preferentially protecting features in the other domain (and so reducing utility of the data in these domains). Developing standard aggregation, suppression or perturbation approaches in each of these domains would assist when analysing data from different sources.

It is certainly possible to imagine standard protection approaches for numerical features (such as latitude and longitude or age in days), but more challenging for categorical features (eye colour or hair colour). An example of how this may be done is given in Chapter 10.

07

Describing safe use of outputs

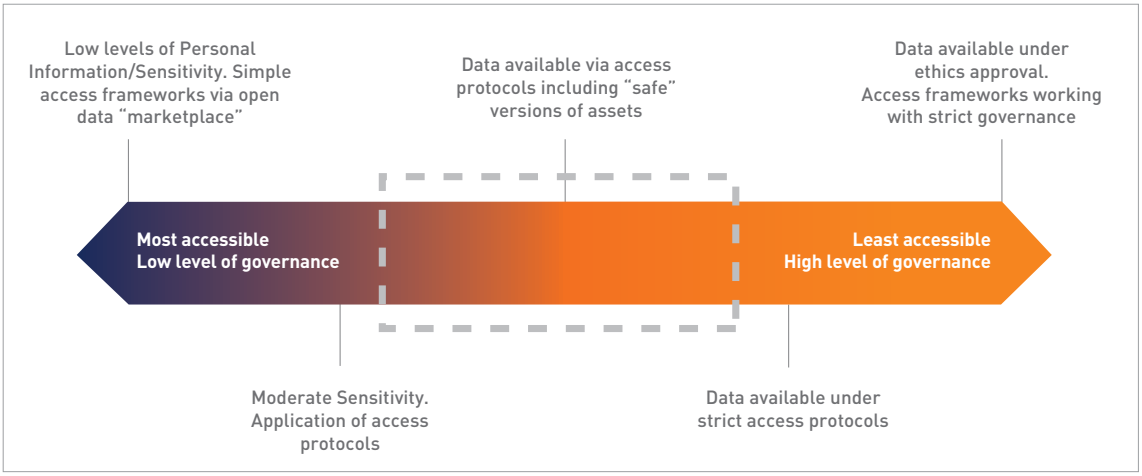


FIGURE 27. CONCEPTUAL SENSITIVITY SCALE

While the restrictions to data sharing are often related to privacy, many other concerns relate to consequences of use of data.

Broadly, these are described as sensitivities of data and will be explored separately from privacy concerns. Figure 27 highlights the need for lower (left-hand side) or higher (right-hand side) levels of governance and support or expert interpretation required for use of data (and production of outputs) of different sensitivity. Figure 28 shows a conceptual framework which allows us to consider the aspects of sensitivity separately from privacy.

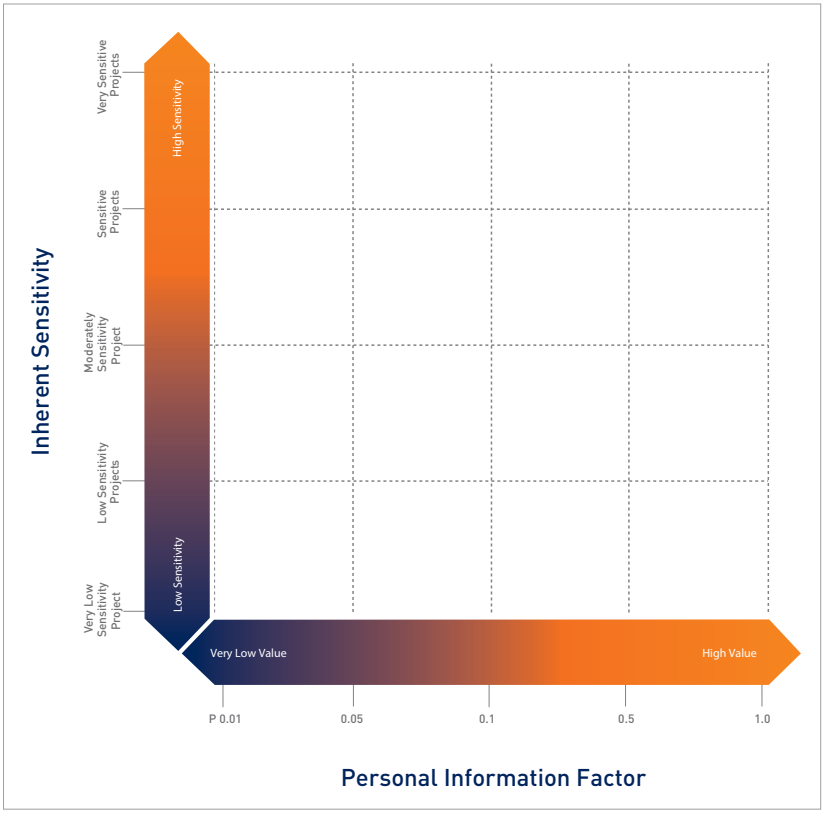


FIGURE 28. SENSITIVITY VERSUS PRIVACY

SENSITIVITY COVERS MULTIPLE DOMAINS, INCLUDING:

- Sensitive subjects captured in data (subjective, but often described in different economies).
- Concerns about the consequences of outputs being used.
- Concerns about the loss of agency (control) for the data holder.
- Unexpected results from analysis, leading to negative surprises or embarrassment.
- Concerns about the ability to appropriately interpret results.
- Concerns about results generated from poor-quality data.
- Concerns about results generated with poor analytical quality.
- Concerns about accidental release of data or results.
- Concerns about data age.

ADDRESSING CONCERNS AROUND USE MAY INCLUDE:

- Frameworks to evaluate consequences of the use of outputs.
- Frameworks to determine the level of confidence in the accuracy of outputs.
- Frameworks to determine the completeness of outputs (how much contextualisation is needed).
- The degree of automation associated with the use of outputs.
- Frameworks to evaluate the "expectedness" of outputs (high or low information content).
- Long-term monitoring of consequences of use of outputs.

Sensitivity example: use of outputs based on context required and unexpectedness of result

Figure 29 illustrates an example framework for "safe use" of data and outputs when considering two dimensions of sensitivity: contextualisation required to interpret an output, versus unexpectedness of an output.

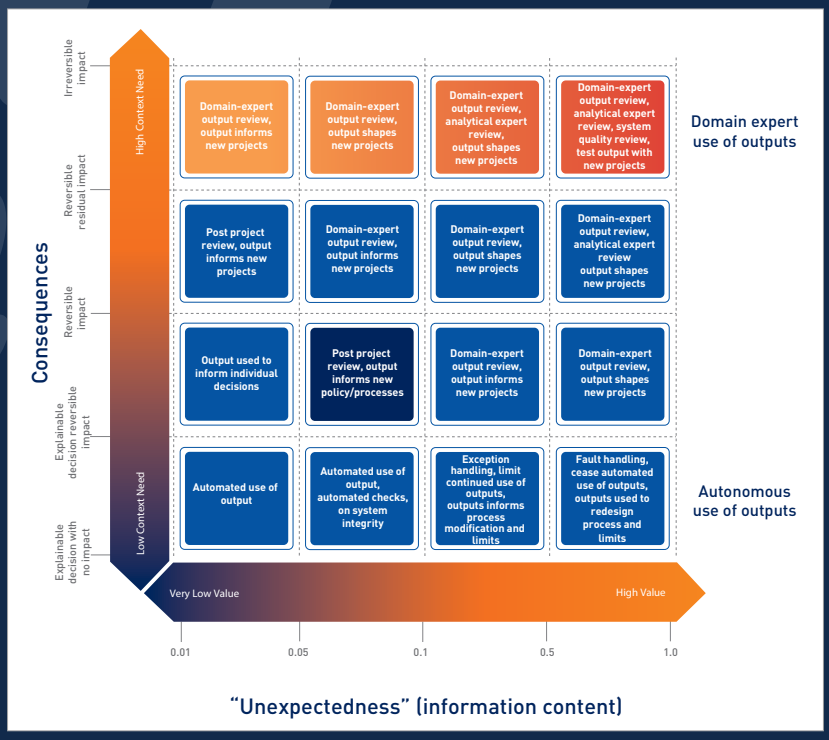


FIGURE 29. APPROPRIATE USE OF DATA AND INSIGHTS — HARM VERSUS UNEXPECTEDNESS

In the bottom left corner, a low-context, expected result may be a trigger for an automated action, such as keeping a train door open for briefly longer than normal period in response to a measured high passenger flow. Operation of the system then continues as normal.

The bottom right may be when this same situation leads to fault-handling, when the same analysis leads to a highly unexpected result outside of normal operating parameters. Automated use of output ceases; the fault requires intervention and use of outputs cannot proceed until the fault is assessed and addressed.

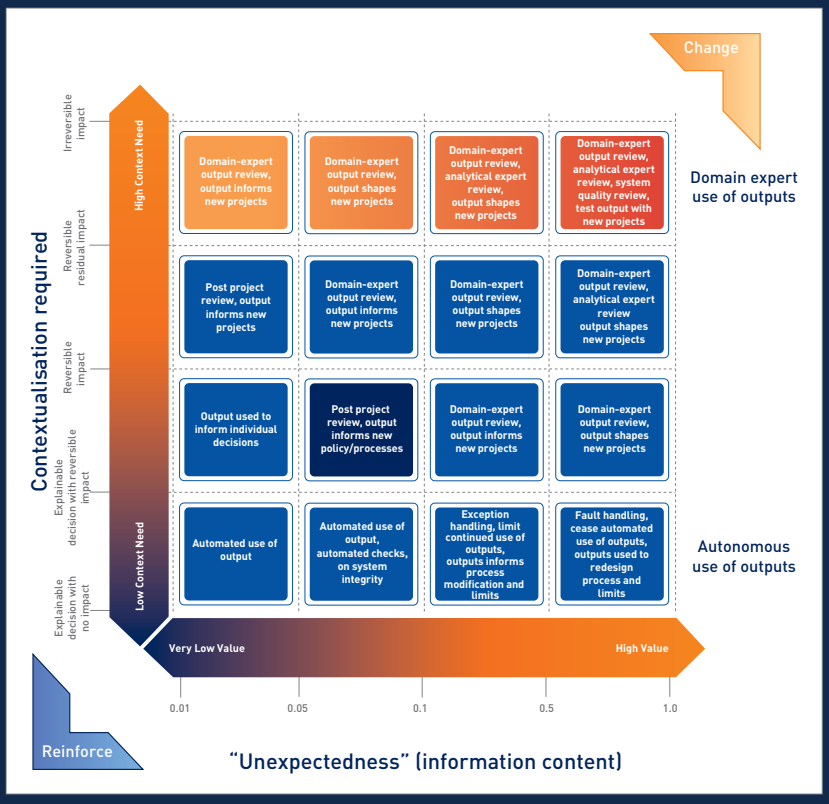


FIGURE 30. APPROPRIATE USE OF DATA AND INSIGHTS — EXPLAINABILITY AND HARM, VERSUS UNEXPECTEDNESS

The top left corner represents a largely expected result in a situation which nonetheless requires expert contextual knowledge, such as a health or human services environment. The output will confirm expected results and inform new projects.

The top right corner is an unexpected output in a high-context environment. Such a result leads to the need to confirm all aspects of the analysis, scrutiny of system and process performance and, ultimately, confirmation of the result through independent analysis. Each square highlights governance controls needed for the combination of sensitivity considerations. Figure 30 shows a similar set of governance considerations for dealing with the sensitivities of unexpectedness of output versus consequences arising from use of the output.

Other safe use considerations

Other factors which drive safe use of outputs include:

- **Explainability -** if outputs are explainable, it is more likely they will be understood and used appropriately.
- **Reversibility -** some decisions are fully reversible simply at an individual's request, others require more effort (up to commencement of litigation); some decisions are only reversible by those with expertise to understand the nature of the decision. Some reversable decisions may still result in residual harms.
- **Decision-maker and legal context -** government decisions are different in qualitative impact to private sector decisions. People can choose (usually) not to deal with a private company, but we are all bound by government decisions. Government decisions also have basic standards they have to meet – rule of law and public law requirements. This may include explainability (see above) but also transparency, accountability and equal treatment. Private sector decisions have lower legal requirements (such as compliance with discrimination law). Decisions that may not satisfy legal requirements (for the relevant sector) would be very high sensitivity.
- **Extent of impact -** at the top of the scale are impacts that affect fundamental rights (such as loss of liberty, children removed), then significant life impacts (admission to university, employment decisions, significant financial impact), then average impacts (moderate financial impact), minor impacts (requirement to participate in something, minor embarrassment) and negligible impacts (don't receive marketing material).
- **Impact on whom -** are the individual(s) impacted in an already vulnerable group?
- **Granularity of impact -** how large is the impacted group? This can work both ways – in terms of spreading but also increasing the impact.
- **Control of impact -** is the impact chosen by a trusted group or is an output released that will allow a broader group to select the impact? Can the impact be controlled? All factors need to be assessed with respect to all potential relevant actors.

08

Safe Data — a starting point

This chapter provides a starting point for using the PIF and Utility measures for different levels of data protection, and a means for dealing with the challenge of trajectories.

The RIG_{95}, RIG_{max}, and PIF are as described in Chapter 6.

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).

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

To understand the relative Utility (μ) of a generated dataset, the MI can be normalised to values between 0 and 1 by dividing it against the mutual information of the original data itself ***I(X; X)***:

$$D_{\mathrm{KL}}(P\|Q) = -\sum_{x\in\mathcal{X}} p(x)\log q(x)$$

A relative Utility of 1 implies no information loss compared to the original dataset. A relative Utility of 0 implies complete information loss in the resultant dataset.

Figure 31 shows an example of relative Utility declining as a feature "age" is aggregated into two-year, then five-year and then ten-year bins in a dataset.

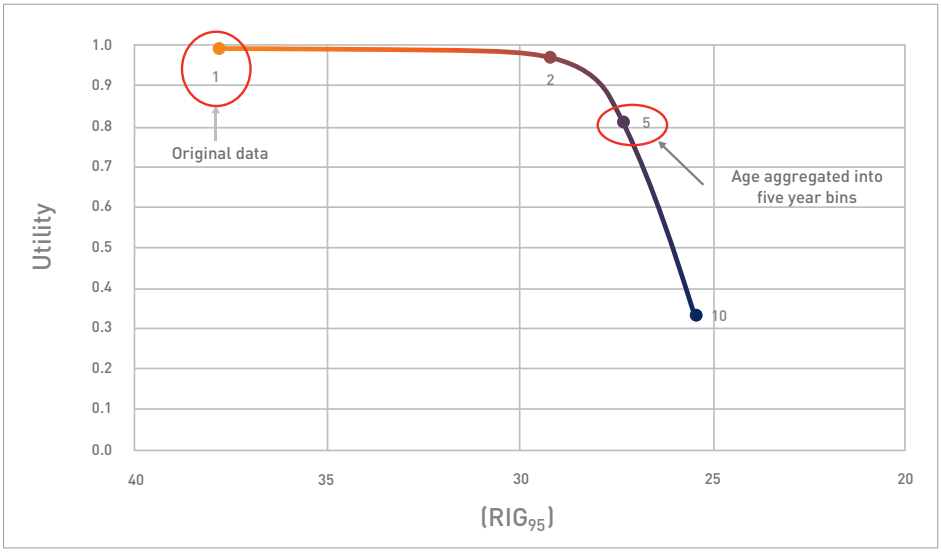


FIGURE 31. EXAMPLE OF DECREASE IN RELATIVE UTILITY AND RIG_{95} AS "AGE" IS AGGREGATED INTO BROADER BINS

Dealing with trajectories – a starting point

Trajectories, pathways or journeys (linked rows in a dataset) represent a very significant weakness when exploring risk of re-identification of an individual with people-centred data. The challenge is that, given the dimensions of personal features, relationships, location in space and location in time, a unique trajectory can very readily be created. Analysis of the example datasets in used in a series of ACS' Directed Ideation series in February 2019 and July 2019 showed that any non-trivial linked trajectory dataset will readily reveal uniqueness and so presents a high risk of re-identification of individuals.

The starting point for dealing with trajectories was based on subsequence decomposition that considered both continuous and non-continuous subsequences of all the features that could be used to create a trajectory.

For example, with the hospital admissions dataset (dataset 7), a trajectory for each patient could be constructed based on time of visit, hospital venue or reason for admission. Each of these individual features, or combinations of features, 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 would be possible to identify a unique trajectory.

The overall approach to solving this is:

- Every individual has a linked set of rows that forms a sequence. Each of these linked rows has a number of features that could form a trajectory in isolation or as groups.
- Find all possible (continuous and non-continuous) subsequences from the main sequence. Group them into 1-step, 2-step, ... N-step sub-sequences.
- For each possible subsequence in the dataset and each number of steps, determine the number of individuals with this subsequence.
- The subsequences with the greatest re-identification risk are those associated with only one individual (a unique trajectory).
- Determine the maximum allowed number of steps that does not lead to a unique trajectory.
- Repeat this process for all features that can form a trajectory.

A worked example for the dataset 1 (Inmate Admissions) is shown in Figure 32. 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* subsequences can be calculated from the number of unique venue transitions (continuous and non-continuous) as shown in Figure 33.

In this example, at subsequence length 1 and 2, there are no unique trajectories for the feature "Admissions".

At length 3, there are 39 unique trajectories, 25 trajectories with 2 identical members (MICS of 2), 24 with 3 (MICS of 3) and so on. At length 4, there are 410 unique trajectories. At length 15, there are 368 unique trajectories.

Based on the PIF analysis described earlier, an approach to making datasets "safer" is to ensure there are no unique trajectories (or set minimum numbers of members in the smallest cohort). This analysis therefore provides a way of testing for uniqueness and for identifying areas for focus to make the data more Safe.

This approach clearly 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. Combinations of features are likely to reduce this below 3 steps to obtain unique trajectories.

id	Admisssion sequence
19	('DE', 'DE')
20	('DE', 'DE', 'CS', 'DE', '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 32. SAMPLE OF INMATE ADMISSIONS DATASET (DATASET 9) WITH TRAJECTORY BASED ON PRISON SITE

No. of subsequences / No. of steps	No. of individuals with subsequence									
	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 subsequence of length 1 or length 2 is associated with only 1 individual

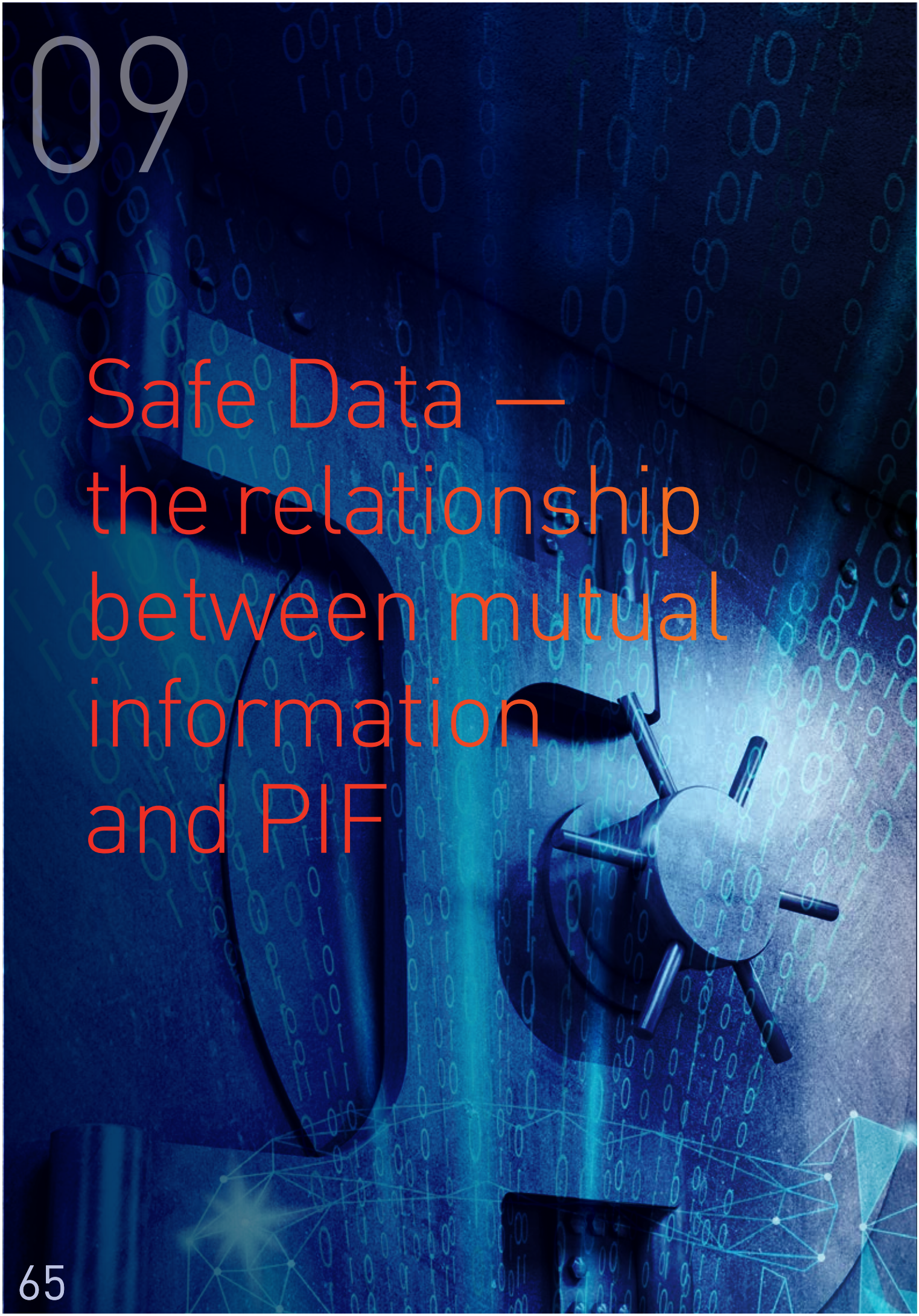
Each of these length-3 subsequences is associated with only 2 individuals

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

FIGURE 33. NUMBER OF SUBSEQUENCES OF DIFFERING LENGTHS FOR INMATE ADMISSIONS DATA (DATASET 9)

09

Safe Data — the relationship between mutual information and PIF



Assessing features in a dataset based on Feature Information Gain

The concept of Cell Information Gain is based on a KL divergence measure of information gained (in bits) by 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 re-identification.

Figure 34 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 date of birth (DOB) are high-risk features from a re-identification 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 re-identification. Gender is seen to be low risk for the entire population, indicating a balance of genders in the dataset population.

As the numerical valued features in the dataset are aggregated, the values of the FIG bands change as shown in Figure 35. The aggregation performed in this example considers every feature to be independent. As features are aggregated, the FIG changes in almost all bands. Nonetheless, the challenge of outliers remains, indicating that further aggregation is required.

	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 34. FIG BANDS FOR FEATURES OF DATASET 6 (SYNTHETIC NAPLAN TEST RESULT DATA)

	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 35. FIG BANDS FOR AGGREGATED FEATURES OF DATASET 6 (SYNTHETIC NAPLAN TEST RESULT DATA)

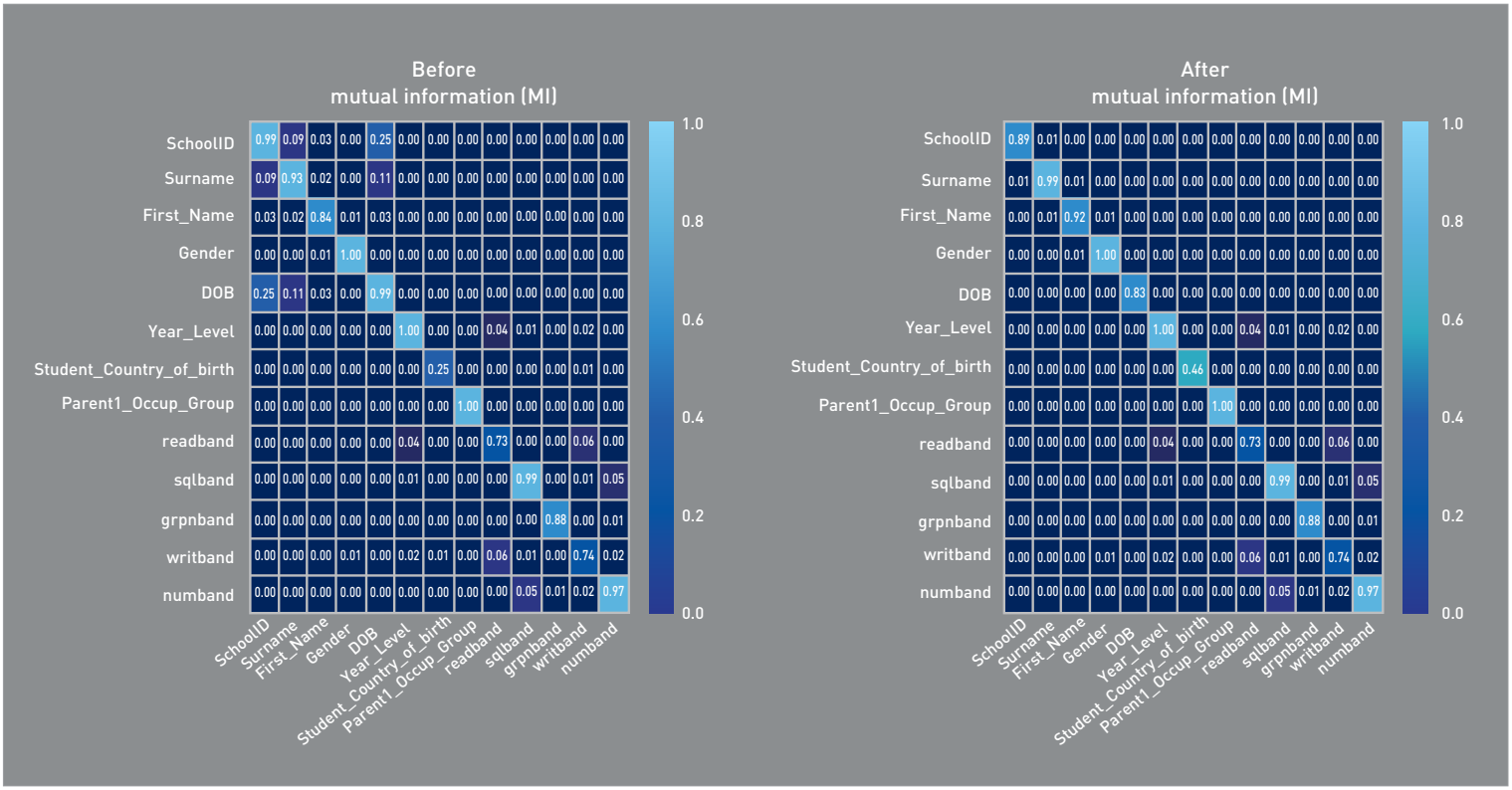


FIGURE 36. MUTUAL INFORMATION BETWEEN FEATURES BEFORE AND AFTER AGGREGATION (DATASET 6)

Feature dependence based on mutual information

In the discussion to date, the ability to infer information between features has been largely ignored for simplicity. It has been assumed that, for example, age and height are independent features. An infant, however, is extremely unlikely to be over 180cm in height, so the features are clearly linked by real-world constraints.

The introduction of the concept of mutual information (MI) allows an exploration of feature dependence and gives insights into which features represented the highest risk of re-identification. 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 36 shows the mutual information between all pairs of features in dataset 6. A high value refers to a high level of mutual information. In the original dataset (on the left), 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_Country_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 (on the right), 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_Country_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 the introduction of more outliers. Not all protection-through-aggregation techniques are the same.

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 9 (NSW Public Service 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, which effectively treats each feature as an independent dataset. Figure 37 shows the change in PIF (see Chapter 6) for each single-feature dataset versus the loss in mutual information between original and aggregated dataset. From this graph, it could be concluded that there is significant reduction in PIF from aggregating "salary", which makes it a more obvious target for protection through aggregation compared to "age" and "years in job".

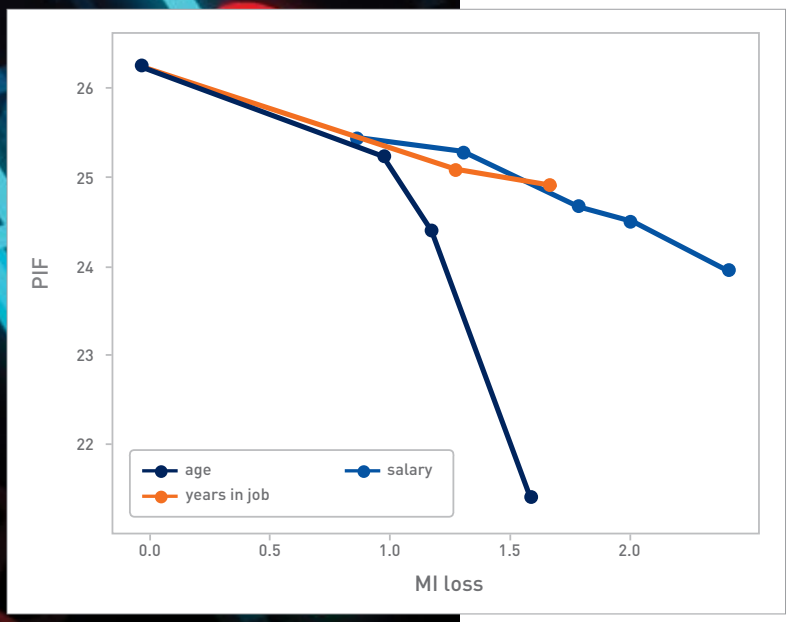


FIGURE 37. A MEASURE OF PIF VERSUS MI OF AGGREGATED DATA FIELDS (DATASET 9)

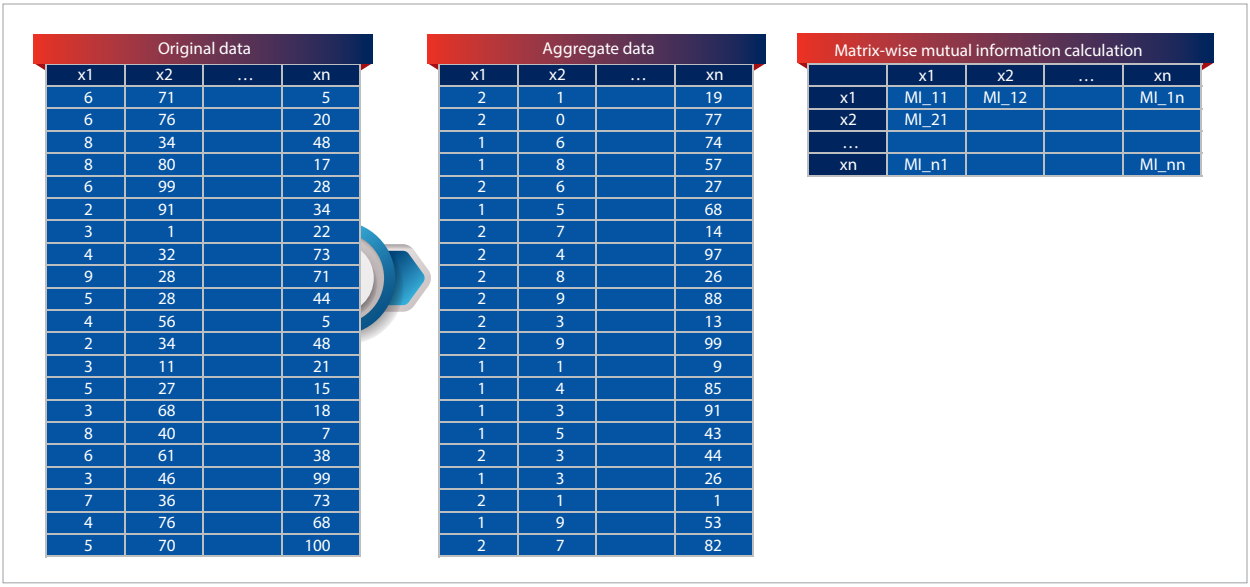


FIGURE 38. MATRIX OF MUTUAL INFORMATION CONCEPT

If, however, dependencies 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) describes the loss between a feature in a dataset and an aggregated version of the same feature (see Figure 38). An MMI allows a more fine-grained analysis of which features to focus on for aggregation.

The approach to using this information is to:

- Calculate the pairwise mutual information between features in the original dataset (as discussed in Chapter 9) to create a mutual information matrix (original MI matrix).
- Aggregate each feature individually to produce a more Safe dataset.
- Calculate the pairwise mutual information between each feature in the original dataset and each feature in the aggregated dataset, the 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 39 outlines this process. Calculating the MMI loss provides a means to track information loss as aggregation is applied to make datasets safer.

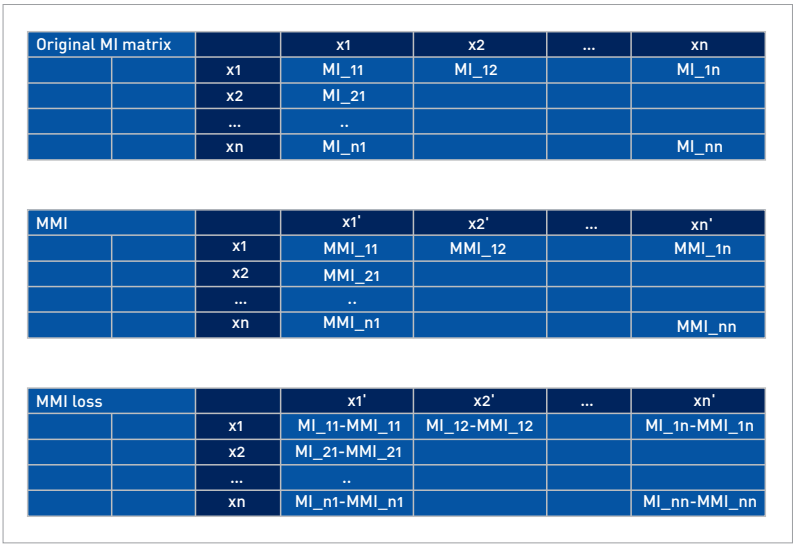


FIGURE 39. MATRIX OF MUTUAL INFORMATION LOSS

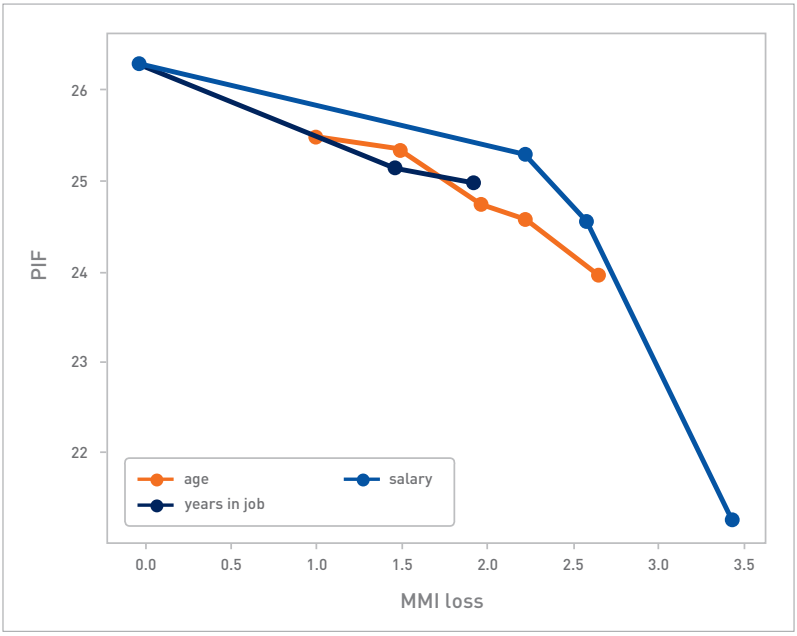


FIGURE 40. MMI LOSS FOR SELECTED FEATURES IN DATASET 8 AS EACH ARE INDEPENDENTLY AGGREGATED

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 40 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, aggregating salary leads to greater MI loss compared to aggregating other features in the dataset.

Implementation

Use of a standard PIF measure and standard thresholds allows the automated production of "safer" versions of a dataset when aggregation (or suppression) are used as the means of reducing risk of re-identification. Figure 41 shows a simple feedback loop that does not consider any specific feature for preferential aggregation. The example method shown is "least two values aggregated", which targets outlier values; however, many variations can be considered.

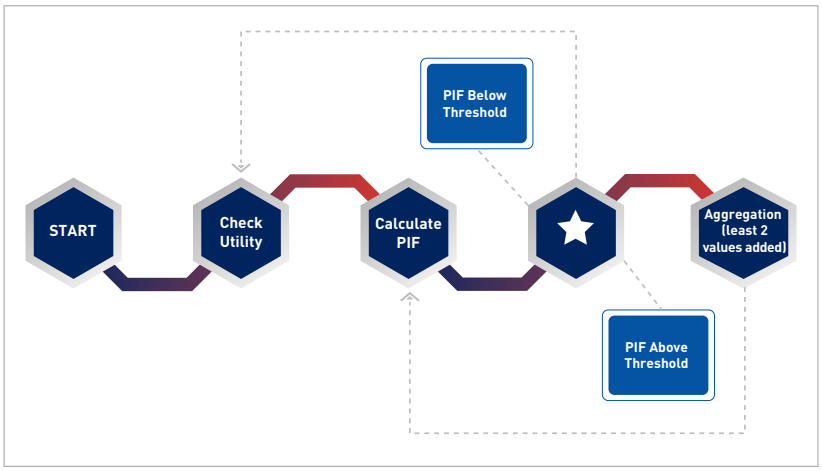


FIGURE 41. EXAMPLE OF AUTOMATED PIF EVALUATION AND REDUCTION BASED ON SIMPLISTIC AGGREGATION

Based on the understanding of the loss of information, an example of a more sophisticated aggregation approach is shown in Figure 42. Many ways of aggregating (or suppression) may be used to protect data, so this should be seen as an example only.

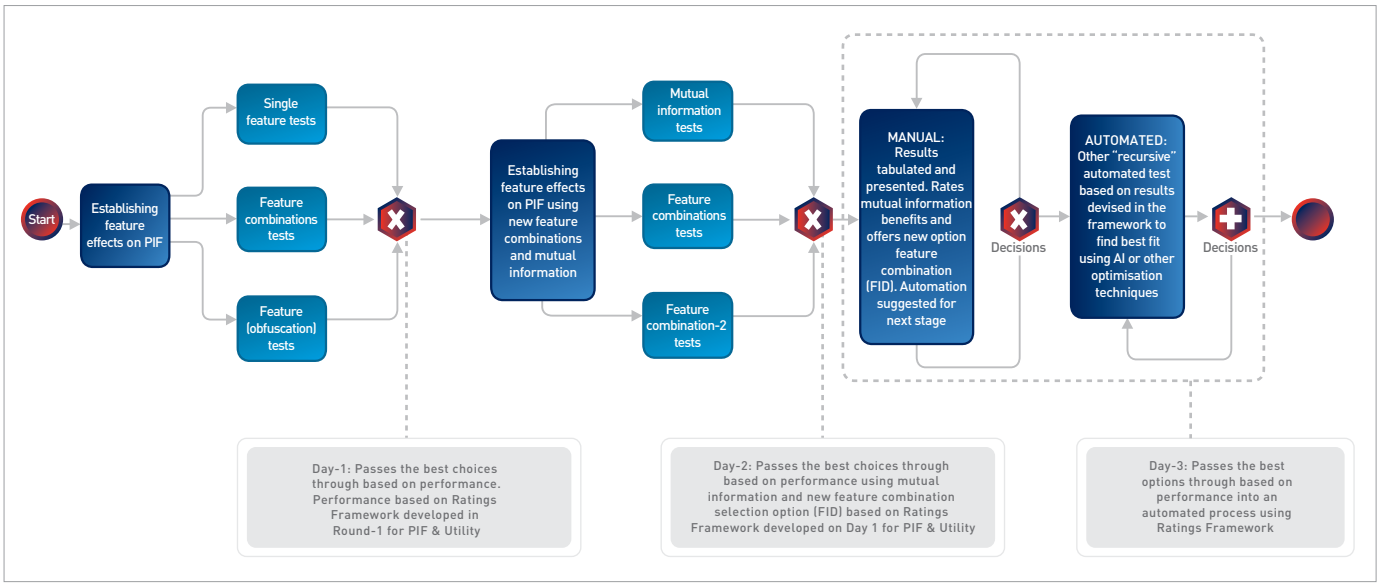


FIGURE 42. EXAMPLE WORKFLOW FOR CREATION OF SAFE DATASETS

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 suppression) of features may also improve dataset Utility.

10

Safe Data — dealing with trajectories

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

Trajectory flattening techniques

The approach described in Chapter 6 was to "flatten" trajectories (see Figure 43) by exploring all possible subsequences for each possible feature (and all combinations of features) that can form a trajectory. The approach can very quickly become computationally intractable as many combinations of subsequence 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.

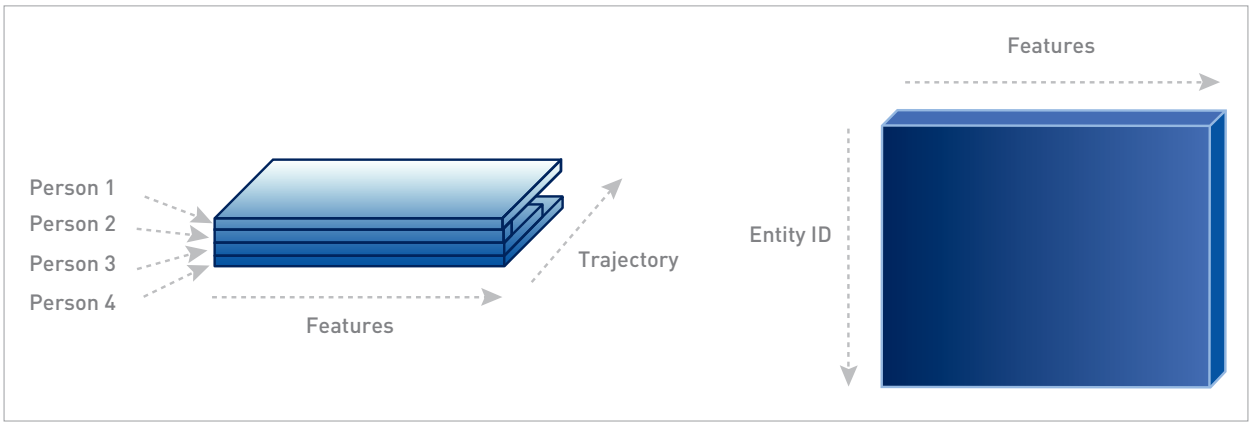


FIGURE 43. TRAJECTORY DECOMPOSITION

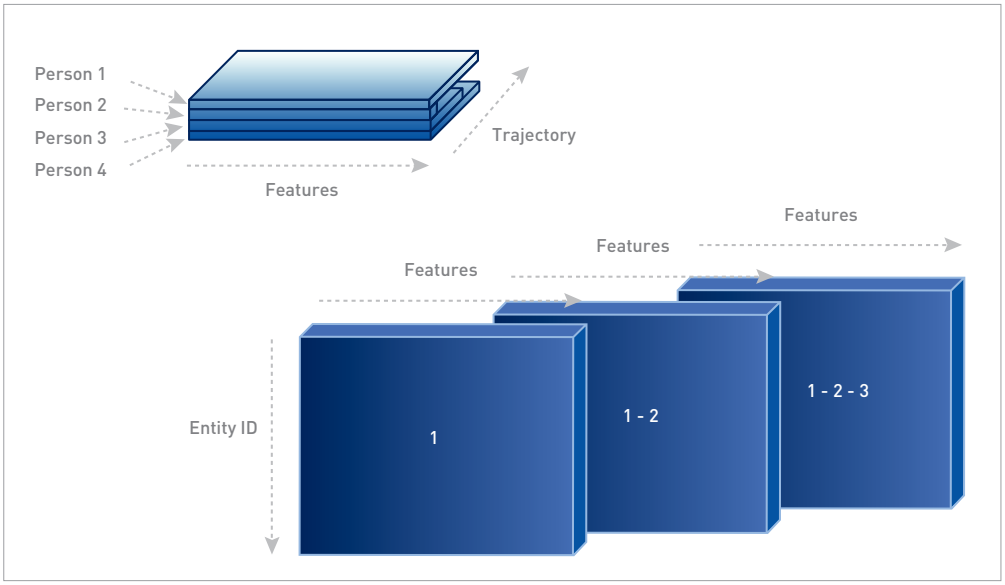


FIGURE 44. DECOMPOSITION OF TRAJECTORY INTO FEATURES

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 for examination, as shown in Figure 44.

An evolution of the subsequence, the DIG approach considers the difference between steps in the sequence and identifies the most unique transitions per stage as shown in Figure 44.

The process makes use of the CIG, which identifies information gain by cell, so that results from the trajectory analysis are comparable to the initial risk identification. The approach is computationally less expensive than identifying unique sequences (and subsequences) and should give a worst-case estimate.

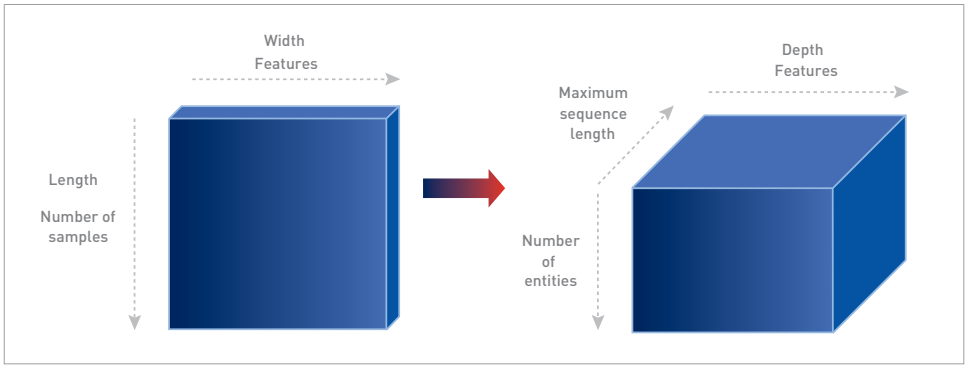


FIGURE 45. TRAJECTORY DECOMPOSITION

LET'S USE DATASET 1 (INMATE ADMISSIONS) AS AN EXAMPLE TO DESCRIBE THE PROCESS (SEE FIGURE 45):

1. Identify a feature that may contain trajectories of interest (such as prisoner ID).
2. Re-shape matrix from sample x feature to I.D. x feature x sample. In the dataset considered, "length" is prisoner ID, "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 ID). For example, if the first three samples for ID 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 ID 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 that may form trajectories.

Step 5 ensures that any given cell will report the highest risk for any sequence it is part of (or its original risk, if that was equal to or higher than any sequence it is part of). Figure 46 shows the evolution of values of the DIG at the first step (DIG baseline) and the final step. Figure 47 shows the corresponding change in mutual information at first step and after completion.

In these tables, the DIG baseline was performed on a subset of the dataset 1, 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.

This approach reduces the complexity compared to full subsequence evaluation. An automated example implementation is shown in Figure 48.

DIG baseline

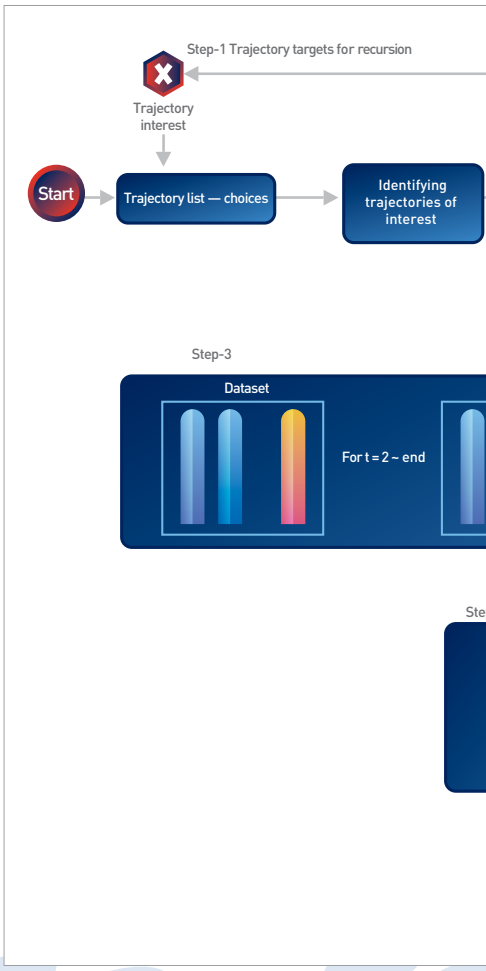
	Min	Q1	Avg	Med	Q3	Max
INMATEID	10.97	13.29	13.22	13.29	13.29	13.29
ADMITTED_DT	11.29	13.29	13.25	13.29	13.29	13.29
DISCHARGED_DT	0.91	0.91	7.06	5.66	13.29	13.29
RACE	0.91	0.91	1.43	1.18	1.18	13.29
GENDER	0.13	0.13	0.74	0.13	0.13	13.29
INMATE_STATUS_CODE	0.41	0.41	1.61	0.41	3.43	13.29
TOP_CHARGE	1.01	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 46. DIG BASELINE STEP 1 (ABOVE) AND AFTER PROCESSING (BELOW) FOR INMATE ADMISSIONS

FIGURE 48. EXAMPLE APPROACH TO DEALING WITH TRAJECTORIES



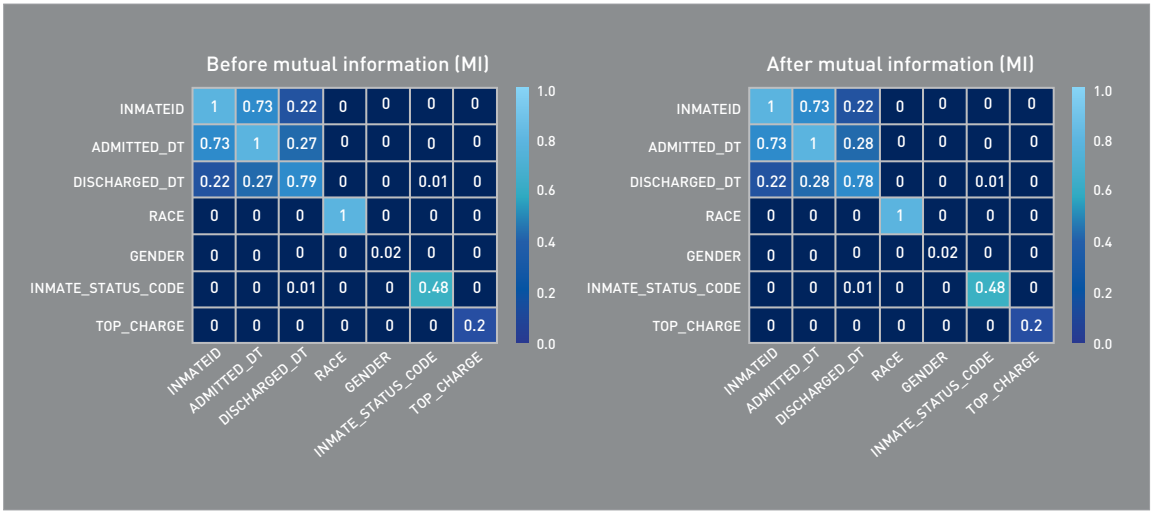
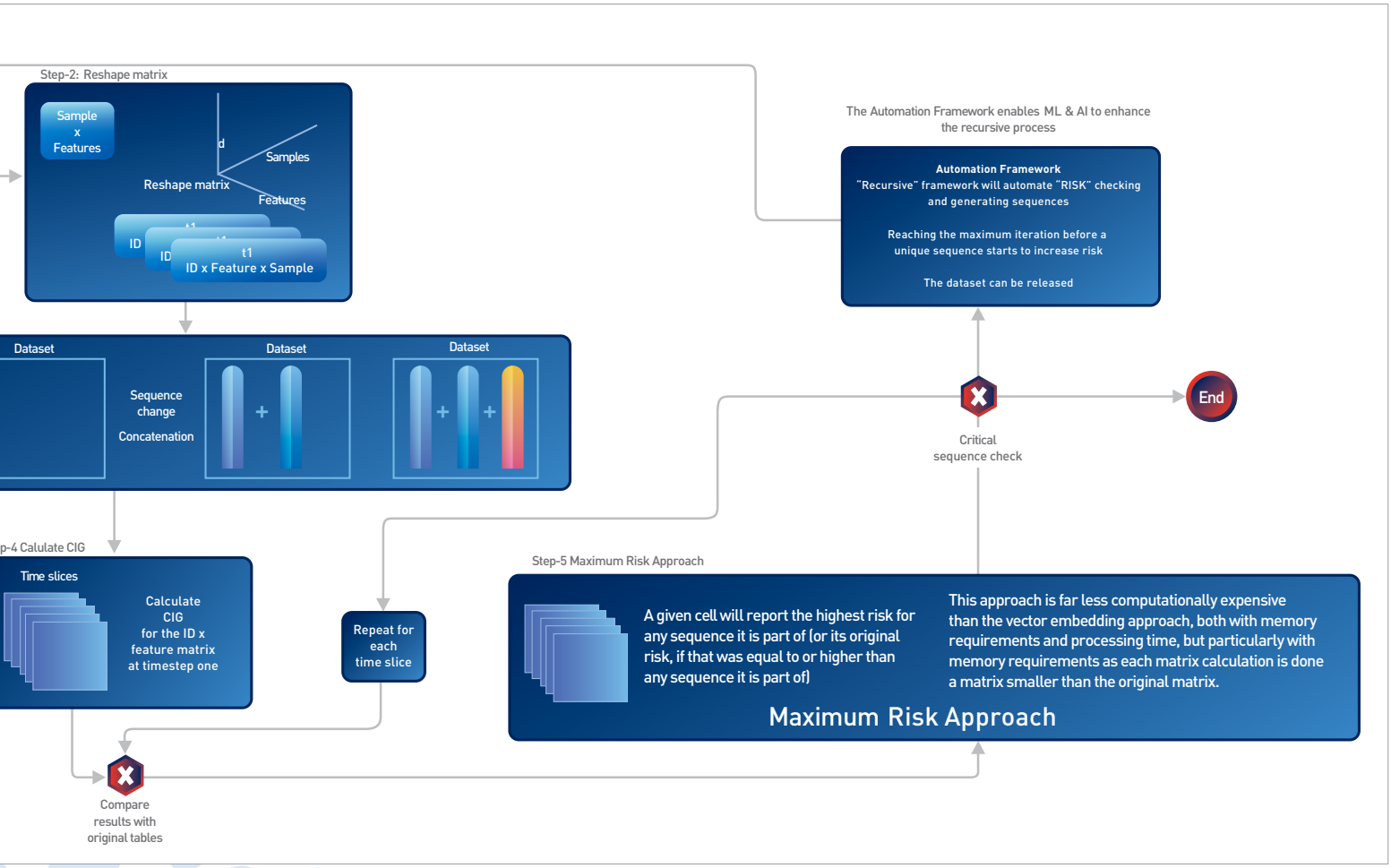
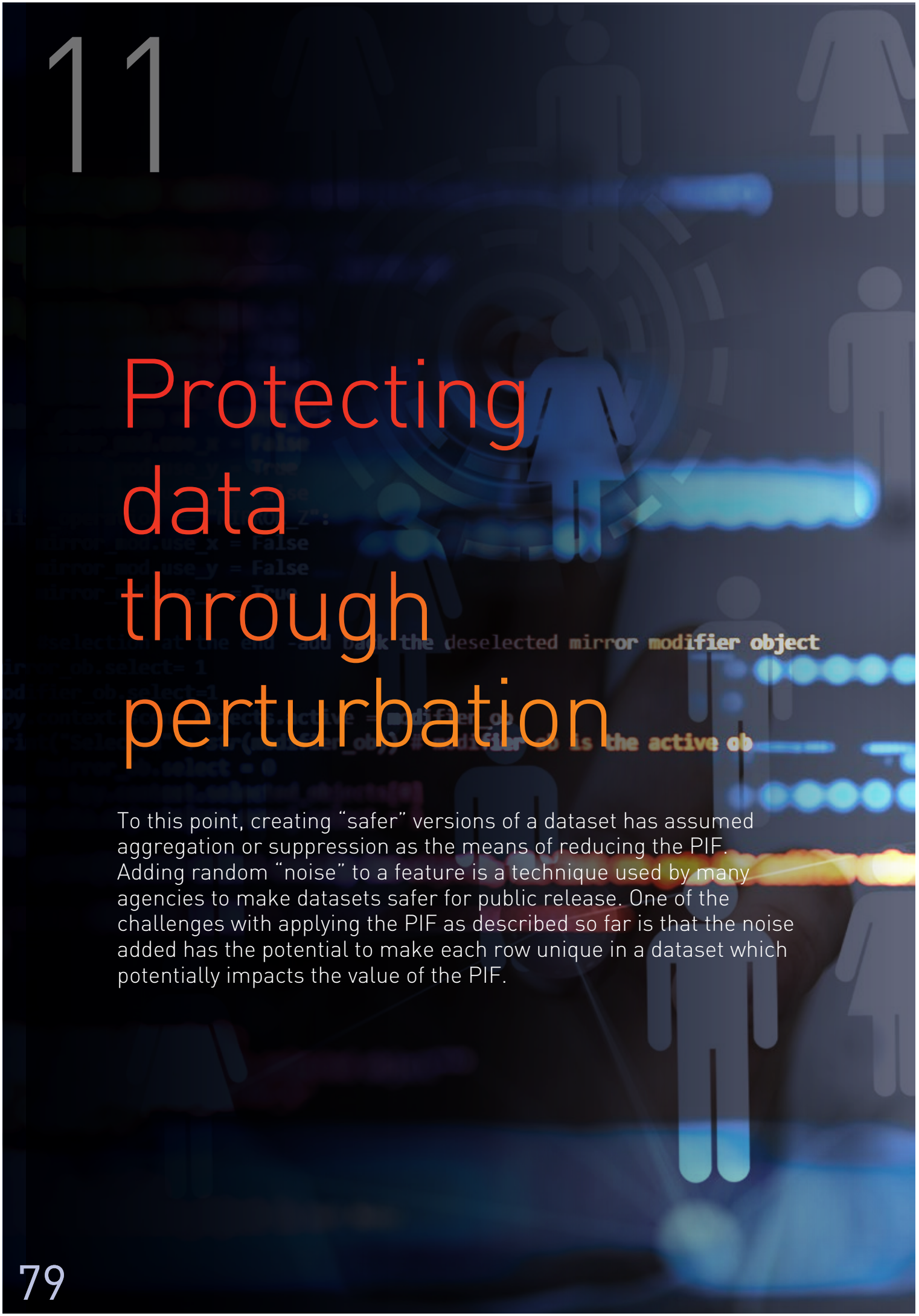


FIGURE 47. MUTUAL INFORMATION AT STEP 1 AND AFTER PROCESSING FOR INMATE ADMISSIONS



11

Protecting data through perturbation



To this point, creating "safer" versions of a dataset has assumed aggregation or suppression as the means of reducing the PIF. Adding random "noise" to a feature is a technique used by many agencies to make datasets safer for public release. One of the challenges with applying the PIF as described so far is that the noise added has the potential to make each row unique in a dataset which potentially impacts the value of the PIF.

Perturbation through random noise is different

Adding random values to a dataset (noise) with a strictly controlled distribution is a common technique for protecting data from the risk of re-identification. Adding noise with a Laplace distribution (see Figure 49) 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.

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 applied 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.[16]

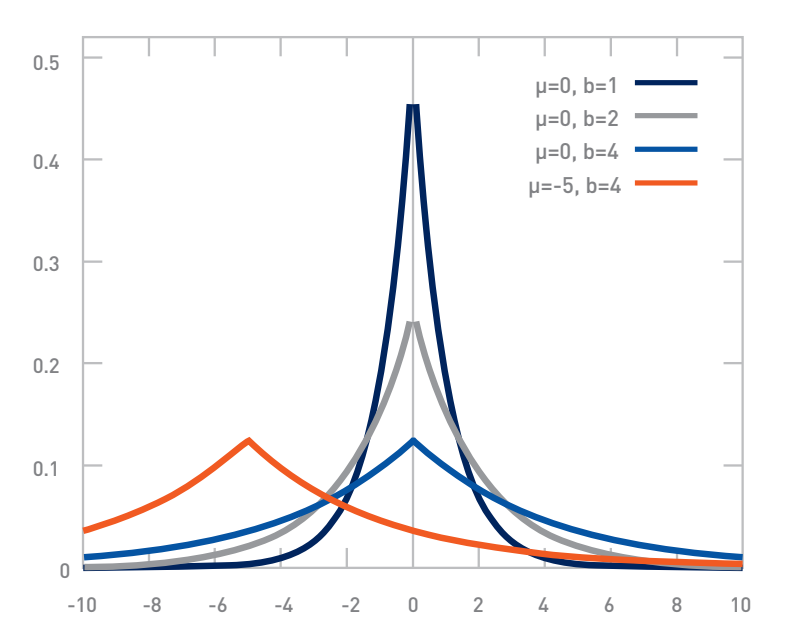


FIGURE 49. LAPLACE DISTRIBUTIONS BASED ON CENTRAL **(μ)** AND DEVIATION **(b)**

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

A differential privacy approach

In recent years, differential privacy has been an active area of research. Differential privacy is a constraint that 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 output of two datasets differing by an individual record, and is represented by the expression:

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

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

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 dataset. 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 dataset. A potential candidate is the MWEM (Exponential Mechanism with the Multiplicative Weights) algorithm (Figure 50).[17] 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 of 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.

[17] 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

Inputs: Dataset 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 $i = 1,\ldots,T$:

1. Exponential Mechanism: Sample a query $q_i \in Q$ using the Exponential Mechanism parametrised 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

$$A_i(x) \propto A_{i-1}(x) \times \exp(q_i(x) \times (m_i - A_{i-1}))/2n).$$

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

FIGURE 50. DESCRIPTION OF THE MWEM ALGORITHM

Figure 51 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 51. DIFFERENTIAL PRIVACY APPROACH

There are a few points about the above approach to note:

- As many executions of MWEM as feasible should be undertaken, and a mean and variance should be computed.
- It could eventuate 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 52.

Generating datasets with different values of epsilon and comparing (see Figure 53) them allows a relative measure of privacy preservation to be explored as shown in Figure 54.

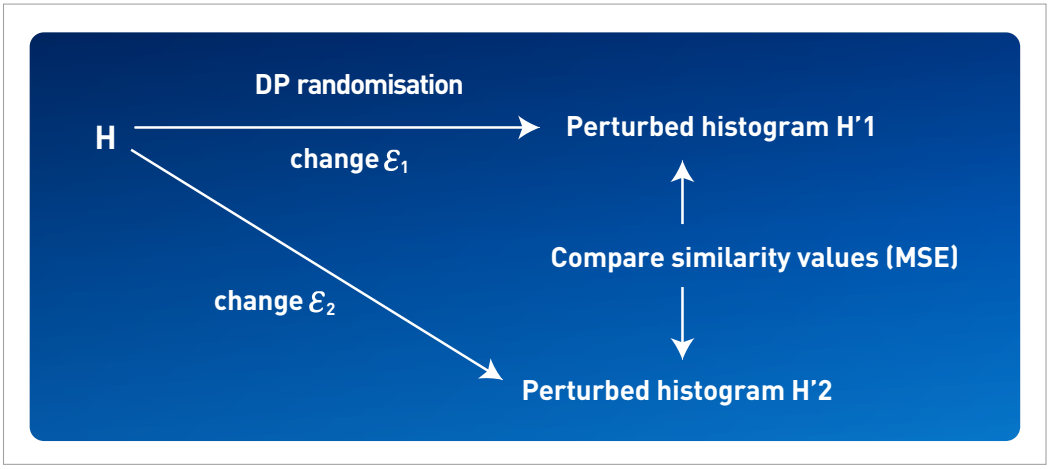


FIGURE 52 COMPARING RELATIVE PRIVACY BETWEEN TWO PERTURBED DATASETS

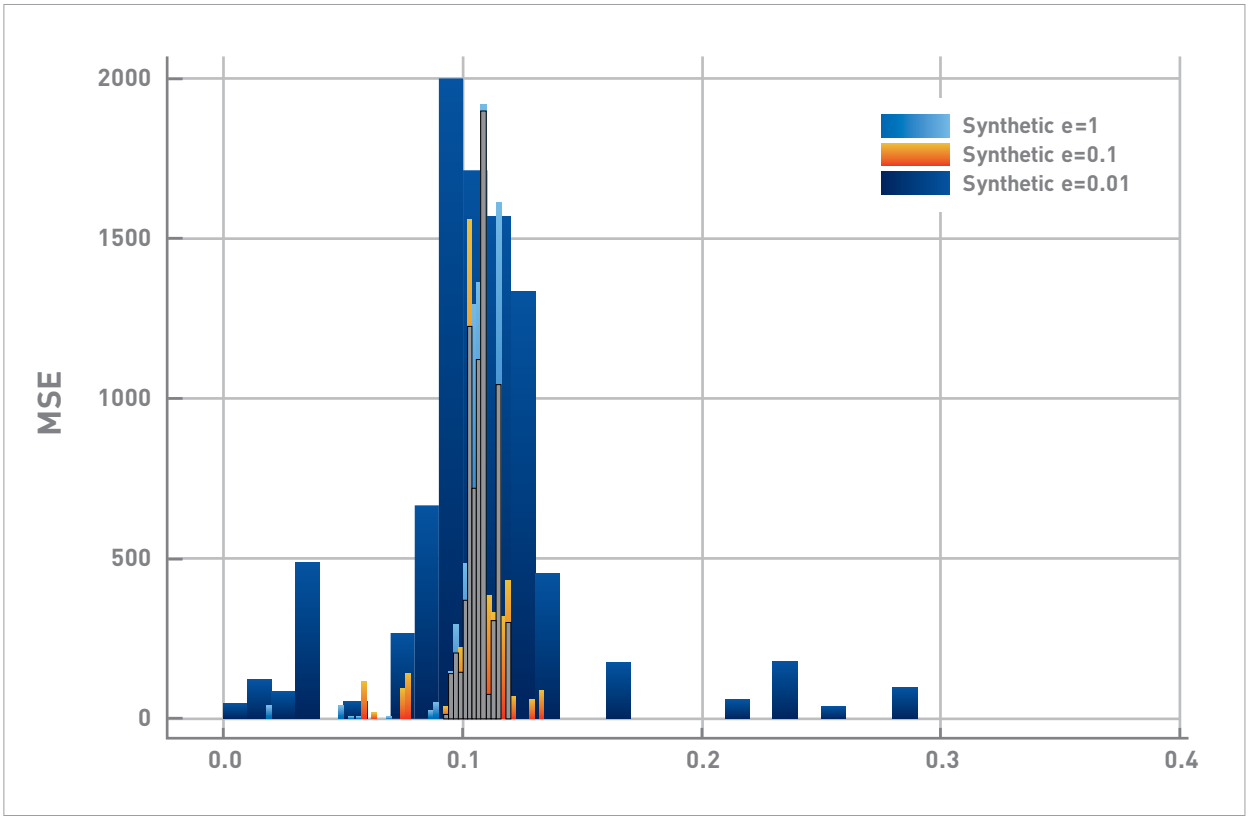


FIGURE 53. SYNTHETIC DATASETS FOR DIFFERENT VALUES OF EPSILON

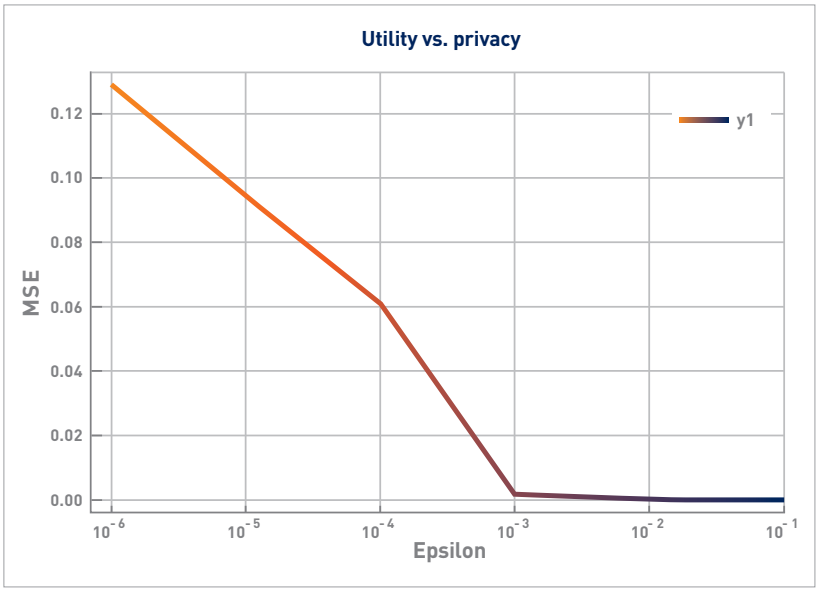


FIGURE 54. SYNTHETIC DATASETS FOR DIFFERENT VALUES OF EPSILON

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.

12

Bringing it all together

Application of controls based on risk

In this chapter, we will bring the pieces together and describe the ways to address the sensitivity versus privacy matrix through controls based on identified risk. After assessing a project for sensitivities, considerations help to address these sensitivities and identify appropriate use of controls, based on the dimensions of the Five Safes and the larger risk framework.

The framework of controls being examined relies on the ability to determine the level of sensitivity of the information captured in the data, the level of PI in the data (PIF). Different controls can be applied at different phases of the project using data (such as collection, analysis, outputs, use of outputs).

The assessment of sensitivities included in these frameworks include consideration of:

- Sensitive subjects captured in data (subjective but often described).
- Consequences of how outputs will be used.
- Results from analysis leading to negative surprises or embarrassment.
- The expert knowledge or context required to appropriately interpret results of analysis.
- Concerns about results generated from poor-quality data.
- Concerns about results generated with poor analytical quality.
- Concerns about accidental release of data or results.
- Concerns about data age (or data which has never been examined).

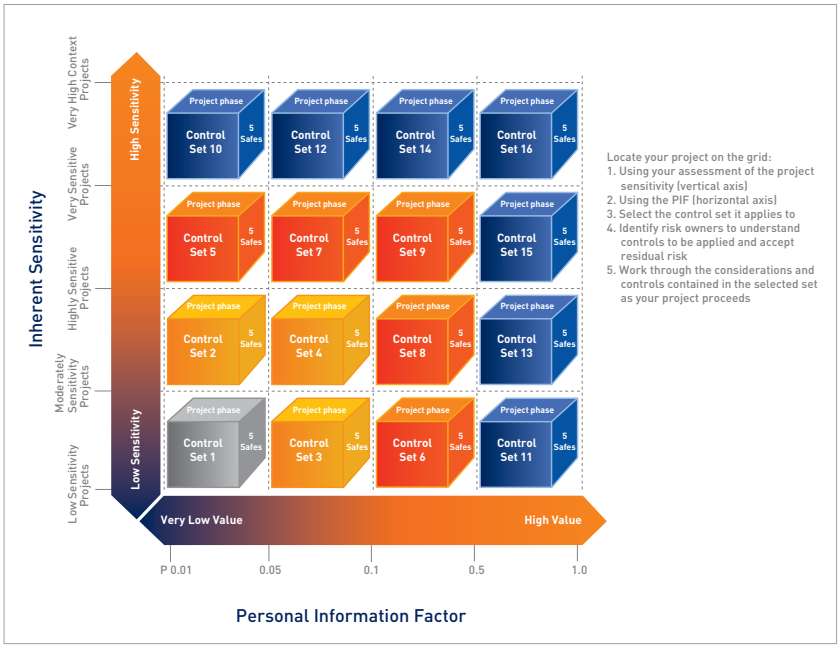
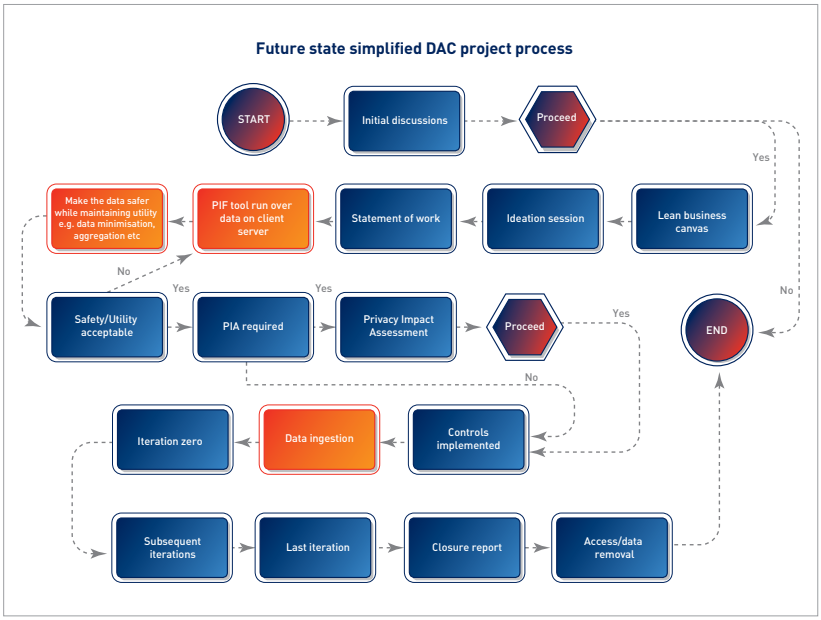


FIGURE 55. CONTROL SET FRAMEWORK

FIGURE 56. EXAMPLE PROJECT FLOW INCLUDING ASSESSMENT OF PIF IN DATA AND SENSITIVITY OF OUTPUTS



Once the sensitivity and privacy of a project is identified, a number of controls can be applied as shown in Figure 56. A decision tree for sensitivity is shown in Figure 57. An example control set for the least sensitivity, lowest PIF projects is shown in Figure 58 (Control Set 1).

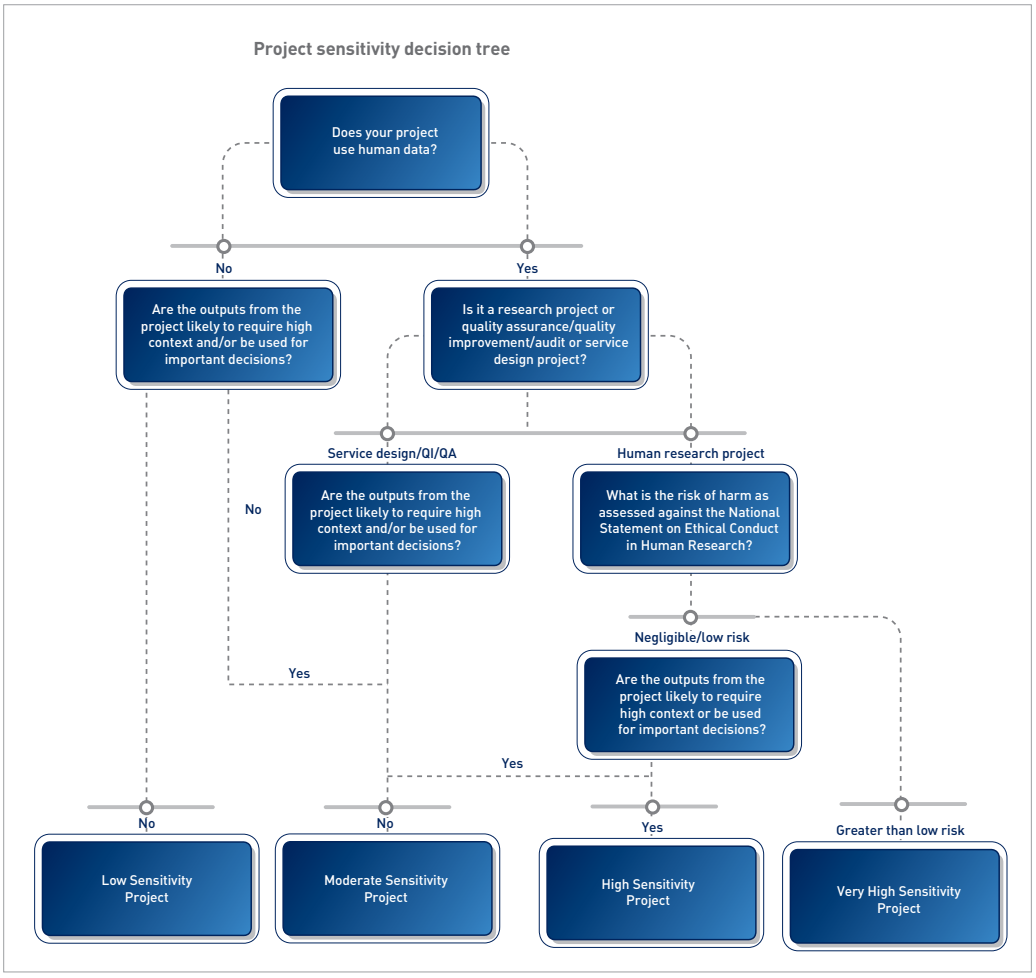


FIGURE 57. EXAMPLE DECISION TREE FOR SENSITIVIITY

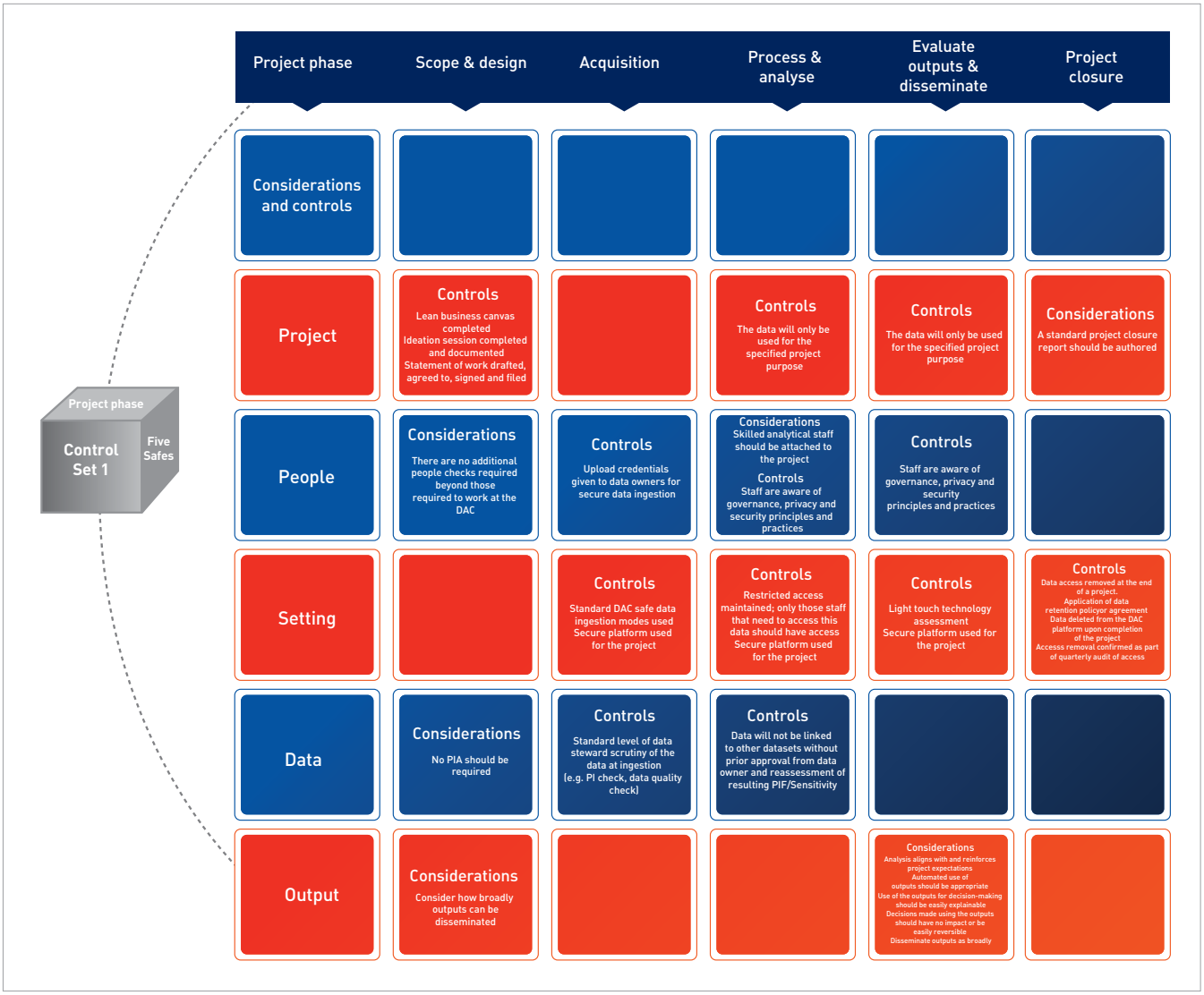


FIGURE 58. EXAMPLE CONTROL SET 1

Examples of trading PIF for Utility using random aggregation

The ability to apply the considerations and controls described above relies on the ability to measure the PIF of data and assess the sensitivity of datasets, and to dial up (or down) controls based on the project's need.

Figure 59 shows an example of arbitrary attempts to aggregate fields in the Inmate Admissions dataset (dataset 1). The bubbles each represent a dataset that is the result of arbitrarily binning each feature into 1, 3 or 5 bins and the impact on the MICS, Utility Factor and RIG_{max}. It should be noted that this is not an exhaustive plot of all possible aggregation combinations and does not represent use of sophisticated aggregation techniques. This figure also does not include the original dataset.

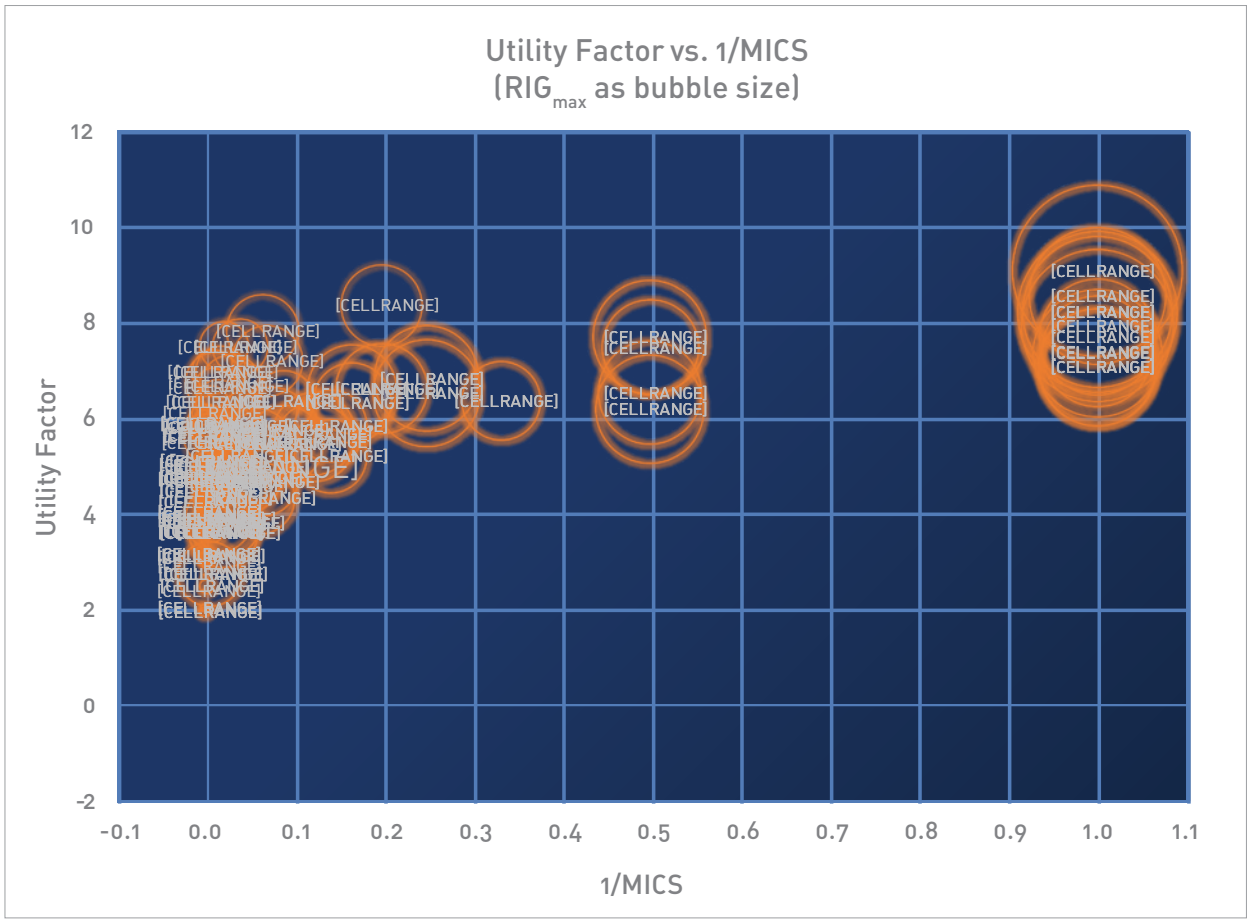


FIGURE 59. UTILITY VERSUS 1/MICS FOR INMATE ADMISSION DATASET (DATASET 1). BUBBLE SIZE IS RIG_{MAX}.

The ideal scenario would be the ability to make a dataset "safer" by reducing the PIF, while maintaining the Utility Factor as high as possible. For a MICS of 1 (unique rows exist), the dataset with the highest RIG_{max} is 2.7 with a Utility of approximately 7.6. When the MICS is increased to 2 (1/MICS is 0.5), the highest RIG_{max} within this sample set is 1.2 with a Utility of approximately 6.4. This gives a PIF of 0.6, which is significantly lower for only a modest decrease in Utility.

From the datasets generated from the forced aggregation of fields, the datasets with the highest Utility are:

MICS 1: RIG_{max} identified 2.7 (Utility 7.6), PIF is 2.7

MICS 2: RIG_{max} identified 1.2 (Utility 6.4), PIF is 0.60

MICS 3: RIG_{max} identified 0.6 (Utility 5.3), PIF is 0.19

MICS 4: $RRIG_{max}$ identified 1.1 (Utility 5.4), PIF is 0.26

MICS 5: RIG_{max} identified 0.8 (Utility 5.5), PIF is 0.15

MICS 6: RIG_{max} identified 0.7 (Utility 5.4), PIF is 0.12

MICS 7: RIG_{max}x identified 0.5 (Utility 4.8), PIF is 0.07

When the MICS is, for example, increased to 5 through random aggregation (1/MICS is 0.2), the largest RIG_{max} value is 0.8, and so the PIF is significantly reduced to 0.15 without substantial loss in Utility.

This example dataset is far "safer" from a risk of re-identification perspective, and still has relatively high Utility. The fact that not every possible dataset is generated means that some datasets generated have a better trade-off of PIF and Utility. Further dataset generation may identify better trade-offs.

Using the example Safe Data thresholds given in Chapter 6, we can now see from the example datasets generated the level of level of aggregation required and the Utility which can be achieved for each PIF threshold:

Safe Level 1: 1.00 ≤ PIF	MICS is 1, Utility is at least 7.6
Safe Level 2: 0.33 ≤ PIF < 1.00	MICS is at most 2, Utility is at least 6.4
Safe Level 3: 0.11 ≤ PIF < 0.33	MICS is at most 3, Utility is at least 5.4
Safe Level 4: 0.04 ≤ PIF < 0.11	MICS is at most 7, Utility is at least 5.3
Safe Level 5: PIF < 0.04	MICS is at most 11, Utility is at least 5.3

Once again, the fact that not every possible dataset is generated means that some datasets generated have better trade-off of MICS and Utility to achieve the PIF threshold required.

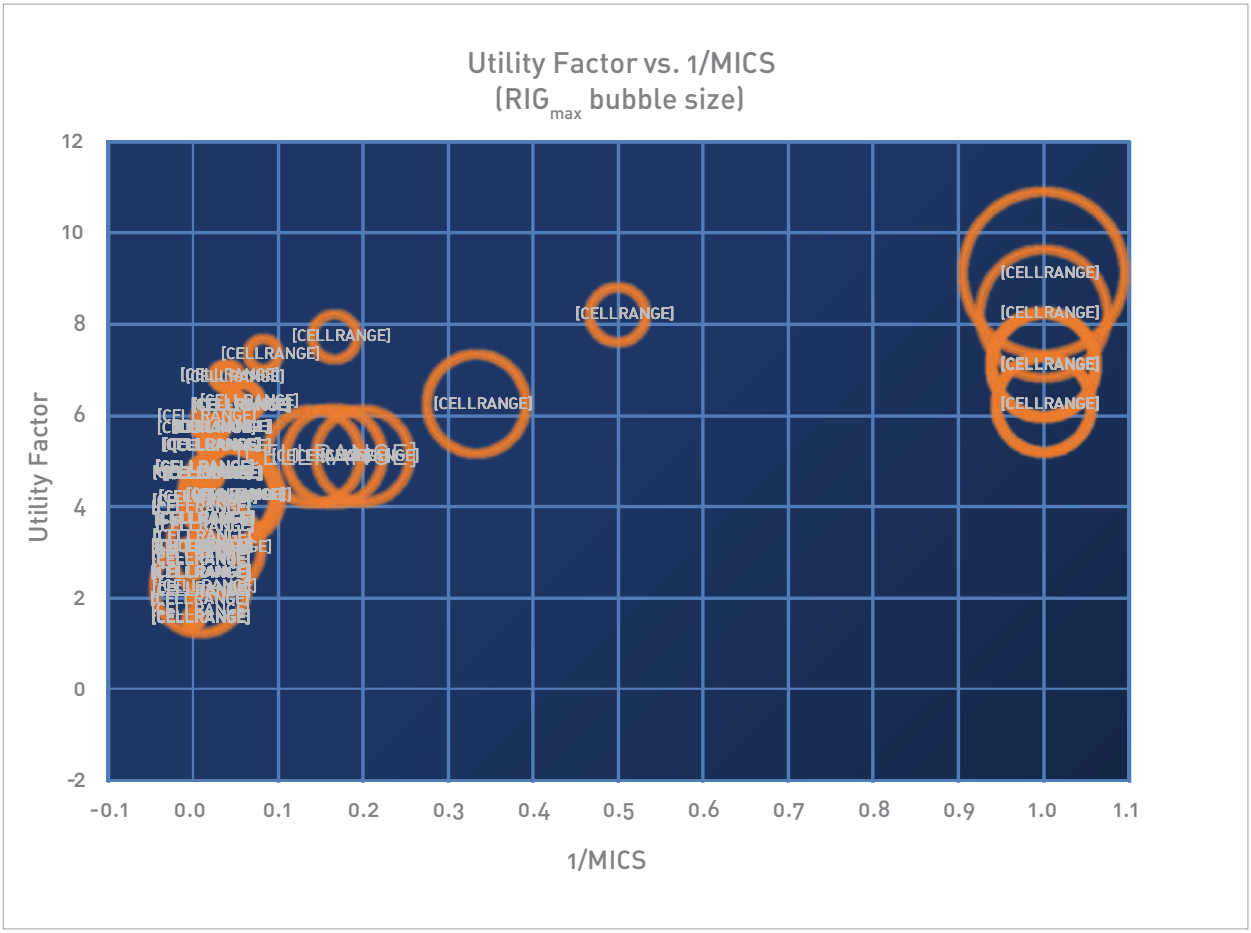


FIGURE 60. UTILITY VERSUS 1/MICS FOR DATASET 9. BUBBLE SIZE IS RIG_{MAX}

Figure 60 shows datasets similarly generated from the Synthetic NSW Workforce Profile dataset (dataset 9) when fields are randomly aggregated into 1, 2 or 4 bins. From the datasets generated from the forced aggregation of fields, the datasets with the highest Utility are:

MICS 1: RIG_{max} identified 7.0 (Utility 9.1), PIF is 7.0

MICS 2: RIG_{max} identified 0.8 (Utility 8.2), PIF is 0.41

MICS 3: RIG_{max} identified 2.6 (Utility 6.3), PIF is 0.87

MICS 4: no sets found

MICS 5: RIG_{max} identified 2.3 (Utility 5.1), PIF is 0.46

The aggregation techniques used are brute force in that an arbitrary number of bins are defined for each field and new datasets are randomly generated with these levels of aggregation. The Synthetic NSW Workforce Profile dataset shows that, even without knowledge of the mutual information between fields, random aggregation can reduce the PIF below a selected threshold while still maintaining moderate levels of Utility compared to the original dataset. The fact than no sets were generated with a MICS of 4 is likely the result of the sample size of sets generated rather than an inherent characteristic of the data itself.

Again, from the example datasets generated, the level of level of aggregation required and the Utility which can be achieved for each PIF threshold:

Safe Level 1: 1.00 ≤ PIF	MICS is 1, Utility is at least 9.1
Safe Level 2: 0.33 ≤ PIF < 1.00	MICS is at most 2, Utility is at least 8.2

Once again, the fact that not every possible dataset is generated means that some datasets generated have better trade-off of MICS and Utility to achieve the PIF threshold required.

13

Discussion

What is a use case and which parts are fixed?

One of the conclusions from ACS' Directed Ideation series was that the intended use of the data was as a very significant factor when determining the risk framework for data. A formal definition of a use case would bring clarity. Very often, however, a use case quickly gets described in terms of many aspects of the Five Safes framework:

A use case is characterised by:

- Who wants to access data?
- Why they want to access data?
- Will they access, change or further share the data?
- What is the level of Utility required of the data?
- What is the PIF of the data itself and of the output of analysis?

The basis of the challenge for a use case is to determine which dimensions of the risk framework are set by the nature of the problem and which need to be adjusted in response to the nature of the problem.

The lake temperature example cited earlier could now be considered in terms of the necessary Utility of the data needed, the PIF inherent in the collected data samples and the PIF appropriate for release of results (to whom and for what purpose). Different results could be generated with decreasing PIF for everwider dissemination including at a PIF suitable for public release.

In terms of the considerations of this project:

- Project (fixed): the merits of the project provides a strong motivation to proceed.
- Data (fixed): the location of sensors near isolated homes means that PIF is likely high.
- People (variable): A high PIF means Very Safe People are needed for the project.
- Setting (variable): A high PIF means a Very Safe Setting is needed for the project.
- Outputs (variable): Project results can be generated at different PIF levels for different users.

Metadata standards

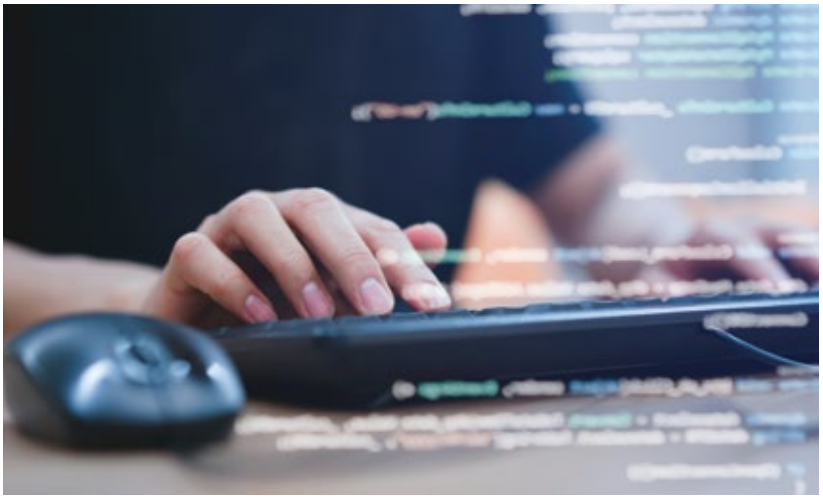
Throughout the Directed Ideation activities underpinning this work, teams used manual or ad-hoc processes to aggregate data to high Safe levels. One team concluded that spatial parameters should be converted to latitude and longitude to give the greatest flexibility for spatial aggregation.

In many cases, standard ways to aggregate common personal, spatial or temporal parameters would increase the level of automation of producing more safe data.

Standard ways of aggregating include:

- Common personal features (such as hair colour, age, eye colour, blood type).
- Common spatial features (location, latitude/longitude).
- Common temporal features (second, minute, hour).

Aggregating or generalising personal, categorical features such as hair colour may be achieved using a process based on standard "identikit" models. To be effective, such an approach must not impose cultural or demographic bias. This is an area for further investigation.



Database reconstruction consideration

Database reconstruction refers to the ability to increase the risk level of re-identification based on repeated and varied access to data, ultimately constructing a granular level of personal information.

If data access is managed by identified users, it may be possible to determine the level of data safety as the number of access requests increases.

Rather than determining data safety per access, taking the aggregate number of and variety of data accesses may lead to threshold conditions for data safety being reached over time. This would then limit further, varied, access or require a user to increase their level of safety. This may lead to a finite number of data products ever being generated to limit reconstruction risk.

13

Conclusions

Throughout this paper we have explored the fundamental challenge of re-identification risk in de-identified data using a Personal Information Factor. An understanding of the degree of personal information in a de-identified dataset, and how effectively different approaches change this level of personal information, should help to create frameworks to balance the risk of re-identification with the Utility of shared data. While our approach is heuristic, the processes we present demonstrate credible ways to consider the challenges of data sharing and – it is hoped – provide a basis for building principles-based data sharing and governance frameworks.

While this paper has focused on personal information, preserving privacy is a cornerstone of any safe data sharing framework. Ultimately, systemising and standardising algorithmic calculations of safe data sharing – with independent verification demonstrating the trust, efficacy and benefit to the community – will be required for Australia to truly benefit from evolving digitally driven developments.

The two Directed Ideation sessions underpinning this work demonstrated the feasibility of measures for personal information based on information theoretic approaches, which work with protection measures based on aggregation and suppression. The work has also demonstrated the feasibility of measures of relative Utility based on mutual information. During the course of these events, improvements were made to protection techniques based on identification of inter-dependence of features in a dataset.

The events 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.

So, while 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 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 re-identification.

Appendix A – 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

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	01/29/2018 03:43:00 PM	BLACK	M	CS	215.5
20191524	01/25/2018 01:33:00 AM	1/08/2018 0:52	BLACK	M	DE	
20190871	1/07/2018 12:20		UNKNOWN	M	DE	
20129999	01/18/2018 11:09:36 AM	01/18/2018 04:12:05 AM	BLACK	F	DE	220.39
20150795	1/12/2018 19:40		UNKNOWN	M	DE	
20178129	01/31/2018 06:05:29 PM	1/09/2018 15:49	UNKNOWN	M	DE	CO
43936	1/09/2018 3:33		UNKNOWN	M	DE	
20191370	01/20/2018 08:03:00 PM	01/26/2018 09:29:01 AM	BLACK	M	DE	265.02
64122	1/05/2018 19:28	1/12/2018 14:20	BLACK	M	CSP	120
165663	1/12/2018 11:50		BLACK	M	DE	
4608	2/06/2016 2:13	1/09/2018 22:45	BLACK	M	DEP	125.25
23108	1/06/2018 14:12	1/06/2018 0:05	BLACK	M	DE	
20190837	1/05/2018 20:08		UNKNOWN	M	DE	

FIGURE 61. SAMPLE OF INMATE ADMISSIONS DATASET (UNITED STATES OPEN DATASET)

Reference:

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

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

Plate	State	License Type	Summons Number	Issue Date	Violation Time	Violation	Judgement 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-MULTILATED/C'FEIT	09/20/2018	65	60	0.5	0.22	125.28	0	112
LASTEVO	NY	PAS	8602692651	5/04/2018	03:19P	FAIL TO DSPLY MUNI METER RECPT		35	0	0	0	35	0	122
21974MG	NY	PAS	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	PAS	8600189032	5/04/2018	08:41A	FAIL TO DISP. MUNI METER RECPT		65	0	0	0	65	0	14
46052MG	NY	PAS	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	PAS	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
744HSG	NJ	PAS	8661602592	05/14/2018	11:49A	NO PARKING-STREET CLEANING		45	0	0	0	45	0	94
46323MG	NY	PAS	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	PAS	8529199674	6/11/2018	01:18P	DOUBLE PARKING		115	0	0	0	115	0	49
46052MG	NY	PAS	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 62. SAMPLE OPEN PARKING AND CAMERA VIOLATIONS (UNITED STATES OPEN DATASET)

Reference:

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

DATASET 3 – AIRBNB 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

27,335 UNIQUE HOST IDs

ID	Name	Host ID	Host Name	Neighbourhood Group	Neighborhood	Latitude	Longitude	Room Type	Price	Minimum Nights	Number of Reviews	Last Reviews	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		Sydeny	-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	2/1/2019	0.05	2	40
15253	Stunning Pe	59850	Morag		Sydney	-3388045	151.21654	Private room	110	2	321	20/04/2019	3.65	3	343
20865	3 BED HOUSE	64282	Fiona		Leichhardt	-33.85907	-33.17275	Entire home/apt	450	7	16	3/1/2019	0.18	1	86
26174	COZY PRIVATE	110561	Amanda		Woolahra	-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 Darlin	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		Waverly	-33.89185	151.27308	Entire home/apt	174	4	128	21/04/2019	1.26	1	140
58056	Studio Yindi	279955	John		Mosman	-33.81927	151.23652	Entire home/apt	140	2	246	8/5/2019	2.41	1	246
58954	Christmas N	282630	Peter		Waverly	-33.89176	151.24259	Entire home/apt	1107	7	0			1	365
61721	2br Eclectic	299170	Eilish		Waverly	-33.8889	151.27726	Entire home/apt	244	4	25	26/02/2019	0.25	1	265
63795	Tree Tops	311659	Tracey		Pittwater	-33.62612	151.33151	Entire home/apt	150	2	63	27/04/2019	0.63	1	306
65126	Large Garde	311659	Nicolette		Waverly	-33.88569	151.26886	Entire home/apt	150	5	11	21/04/2019	0.11	1	40
65635	Russell Hut	318390	Russell		Lane Cove	-33.81079	151.16072	Private room	54	1	165	19/04/2019	1.62	7	81
65857	Private Cou	320878	Jennifer		Sydney	-33.90396	151.19124	Private room	74	2	111	27/04/2019	2.81	1	7
66009	Comfort &	322887	Belinda		Woolahra	-33.88327	151.2275	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	Alittle bit	333581	Brian		Hoensby	-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 63. SAMPLE OF AIRBNB SYDNEY LISTINGS (COMMERCIAL OPEN DATASET)

Reference:

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

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, itemised fares, rate types, payment types, and driver-reported passenger counts.

8.81m ROWS AND **19** COLUMNS

Vendor ID	lpep pickup datetime	lpep dropoff datetime	store and fwd flag	Rate code ID	PULocation ID	DOLocation ID	Passenger Count	Trip Distance	Fare Amount	Extra	Mta Tax	Tip Amount	Tolls Amount	Ehail Free	Improvement Surcharge	Total Amount	Payment Type	Trip Type
2	1/1/2018 0:18	1/1/2018 0:24	N	1	236	236	5	0.7	6	0.5	0.5	0	0		0.3	7.3	2	1
2	1/1/2018 0:30	1/1/2018 0:46	N	1	43	42	5	3.5	14.5	0.5	0.5	0	0		0.3	15.8	2	1
2	1/1/2018 0:07	1/1/2018 0:19	N	1	74	152	1	2.14	10	0.5	0.5	0	0		0.3	11.3	2	1
2	1/1/2018 0:32	1/1/2018 0:33	N	1	255	255	1	0.03	-3	-0.5	-0.5	0	0		-0.3	-4.3	3	1
2	1/1/2018 0:32	1/1/2018 0:33	N	1	255	255	1	0.03	3	0.5	0.5	0	0		0.3	4.3	2	1
2	1/1/2018 0:38	1/1/2018 1:08	N	1	255	161	1	5.63	21	0.5	0.5	0	0		0.3	22.3	2	1
2	1/1/2018 0:18	1/1/2018 0:28	N	1	189	65	5	1.71	8.5	0.5	0.5	0	0		0.3	9.8	2	1
2	1/1/2018 0:38	1/1/2018 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	1/1/2018 0:05	1/1/2018 0:18	N	1	129	82	1	1.61	10	0.5	0.5	0	0		0.3	11.3	2	1
2	1/1/2018 0:35	1/1/2018 0:42	N	1	226	7	1	1.87	7.5	0.5	0.5	0	0		0.3	8.8	2	1
2	1/1/2018 0:21	1/1/2018 0:39	N	1	145	129	2	4.12	16.5	0.5	0.5	0	0		0.3	17.8	2	1
2	1/1/2018 0:56	1/1/2018 1:04	N	1	7	223	2	1.22	7	0.5	0.5	0	0		0.3	8.3	2	1
2	1/1/2018 0:11	1/1/2018 0:30	N	1	255	189	1	4.67	17	0.5	0.5	0	0		0.3	18.3	2	1
2	1/1/2018 0:57	1/1/2018 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	1/1/2018 0:36	1/1/2018 0:51	N	1	244	75	2	6.01	19	0.5	0.5	4	0		0.3	24.3	1	1
1	1/1/2018 0:07	1/1/2018 0:15	N	1	225	37	1	1.9	8	0.5	0.5	3	0		0.3	12.3	1	1
1	1/1/2018 0:25	1/1/2018 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	1/1/2018 0:42	1/1/2018 1:00	N	1	145	173	1	6.9	22	0.5	0.5	0	0		0.3	23.3	1	1
2	1/1/2018 0:06	1/1/2018 0:08	N	1	49	49	1	0.3	3.5	0.5	0.5	0	0		0.3	4.8	2	1
2	1/1/2018 0:34	1/1/2018 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	1/1/2018 0:25	1/1/2018 0:28	N	1	179	7	1	0.5	4.5	0.5	0.5	0	0		0.3	5.8	1	1
2	1/1/2018 0:36	1/1/2018 0:51	N	1	7	193	1	1.82	9	0.5	0.5	0	0		0.3	10.3	1	1
2	1/1/2018 0:53	1/1/2018 1:56	N	1	98	74	1	11.79	36	0.5	0.5	7.46	0		0.3	46.71	1	1
1	1/1/2018 0:11	1/1/2018 0:22	N	1	255	112	1	1.9	9	0.5	0.5	3.05	0		0.3	13.35	1	1
1	1/1/2018 0:40	1/1/2018 1:01	N	1	255	28	1	10.3	29	0.5	0.5	5	0		0.3	35.3	1	1
2	1/1/2018 0:15	1/1/2018 0:25	N	1	80	80	1	1.66	8.5	0.5	0.5	1.96	0		0.3	11.78	1	1
2	1/1/2018 0:35	1/1/2018 0:48	N	1	255	232	1	2.91	12	0.5	0.5	3.32	0		0.3	16.62	1	1
2	1/1/2018 0:55	1/1/2018 1:28	N	1	256	50	1	6.09	25	0.5	0.5	6.58	0		0.3	32.88	1	1
2	1/1/2018 0:41	1/1/2018 0:56	N	1	179	75	5	5.3	17	0.5	0.5	3.66	0		0.3	21.96	1	1
2	1/1/2018 0:36	1/1/2018 0:44	N	1	41	75	1	1.63	8	0.5	0.5	1.86	0		0.3	11.16	1	1
2	1/1/2018 0:48	1/1/2018 0:51	N	1	75	74	1	0.91	4.5	0.5	0.5	0	0		0.3	5.8	2	1
2	1/1/2018 0:56	1/1/2018 1:00	N	1	74	74	2	0.92	5	0.5	0.5	0	0		0.3	6.3	1	1
2	1/1/2018 0:27	1/1/2018 0:34	N	1	7	223	1	0.98	6	0.5	0.5	0	0		0.3	7.3	2	1
2	1/1/2018 0:41	1/1/2018 0:52	N	1	179	7	1	1.42	9	0.5	0.5	0	0		0.3	10.3	2	1

FIGURE 64. SAMPLE OF NYC GREEN TAXI TRIP DATA (UNITED STATES OPEN DATASET)

Reference:

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

DATASET 5 – ATO TAXATION INDIVIDUAL STATISTICS (AUSTRALIAN OPEN DATASETS)

Aggregated individual taxation statistics by industry consisting of 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

Financial Year	Broad Industry 1,4,5	Fine Industry 1	Number of individuals	Taxable income or loss 3no.	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 Vegetables 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	72,394,525	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 65. SAMPLE OF ATO TAXATION INDIVIDUAL STATISTICS (AUSTRALIAN OPEN DATASETS)

Reference:

- FY 2013-14: 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

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, it adheres to the major statistical properties of the original dataset.

School ID	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	7/1/2007	7	1101	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	1/8/2006	7	1101	4	6	5	6	6	8
2417	Cole	Sabrina	2	1/9/2010	3	1101	4	4	6	6	6	5
872	Scott	James	1	3/4/2011	3	1101	2	4	5	3	5	5
1405	Scott	Cheryl	2	7/5/2009	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	12/9/2006	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	9/4/2011	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	5/6/2007	7	1101	2	6	6	6	5	5
1225	Gould	Nicole	2	8/3/2005	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 66. SAMPLE OF SYNTHETIC NAPLAN TEST RESULT DATA (SYNTHETIC DATASET)

Reference:

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

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).[18] 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 AND **14** COLUMNS

96,724 UNIQUE SYNTHETIC PATIENT IDs

Name	Gender	Patient ID	Birthdate	Country of Birth	Address	Blood_type	Eye_color	Job	Company	Visit Time	Age	Diagnosis_ code	Diagnosis_desc
Catherine Phillips	F	287-86-8304	09/09/1928	Jordan	14 Cline Gate	AB+	Blue	Microbiology	Donaldson	1/01/1952 01:35	23	515	Asthma
Thomas Jones	M	533-49-6215	25/06/1933	Holy See (Vatican City State)	1 / 51 Michael	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 Jilian	B-	Brown	NA	NA	1/01/1952 6:30	0	328	Upper respiratory infections
Thomas Petty	M	419-64-4893	12/9/1948	Saint Lucia	34/93 Taylor	AB-	Hazel	NA	NA	1/01/1952 13:55	3	681	Caries of decidous teeth
Mathew Phillips	M	831-18-8881	28/07/1930	Thailand	Level 0 253	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 Hart	O+	Grey	Student	NA	1/1/1952 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	4/8/1942	Bulgaria	Flat 32 580 Eliz	B-	Brown	Student	NA	1/01/1952 18:01	9	681	Caries of decidous 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	A+	Blue	Student	NA	1/01/1952 18:28	7	682	Caires of permanent teeth
Tristan Fisher	M	543-13-6020	28/03/1927	Nepal	Suite 640 9	AB-	Blue	Illustator	Hampton	1/01/1952 18:40	24	547	Migraine
William Garcia	M	664-65-6728	15/12/1941	French Southern Terretories	373 Wilson Ra	O+	Brown	Student	NA	1/01/1952 19:09	10	682	Caires of permanent teeth
Derek Glass	M	637-38-4799	9/3/1939	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	6/6/1934	Vanuatan	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	Leberia	616 Jackson Hil	O-	Blue	Estate age	Davis	1/01/1952 23:06	28	659	Fungal skin diseases
Jonathan Silva	M	580-87-1961	3/8/1936	Antigua and Barbuda	9 Mendoza Ave	A-	Grey	Arboriculture	Reyes	1/01/1952 23:32	28	682	Caries of permanent teeth
James Rivera	M	418-17-8845	1/7/1942	Algeria	Unit 36 316 De	AB-	Brown	Student	NA	1/1/1953 0:22	10	389	Vitamin A deficiency
Cindy Chang	F	068-11-4230	5/12/1936	Italy	1 / 45 Daniel	O-	Hazel	Student	NA	1/1/1953 1:25	16	685	Other oral disorders
Michael Brown	M	464-95-8653	25/06/1946	Finland	6 Howe Terrace	AB+	Brown	Student	NA	1/1/1953 8:25	6	681	Caries of decidous teeth
Ashley Reyes	F	647-43-3234	6/8/1924	Honduras	7 Marks Nook	AB+	Hazel	Scientist	Le,Brown	1/1/1953 10:48	28	548	Tension-type headache
Robert Fuller	M	112-66-3822	16/05/1949	Benin	94 Jill Corso	B+	Brown	NA	NA	1/1/1953 12:22	3	838	Sickle cell trait
Caitlin Ramirez	F	760-48-3377	7/10/1938	Sudan	40 / 674 Alvara	AB+	Grey	Student	NA	1/1/1953 13:18	14	668	Other skin and subcutaneous diseases
Carlos Foster	M	847-54-1496	1/12/1934	Vanuatu	Apt. 259 268	A-	Hazel	Student	NA	1/1/1953 14:05	18	571	Anxiety disorders
Melanie York	F	015-18-7147	1/2/1924	Panama	837 Leonard	A-	Brown	Museum	Faulkner	1/1/1953 14:42	28	609	Premenstrual syndrome

FIGURE 67. SAMPLE OF SYNTHETIC HOSPITAL ADMISSIONS DATA (SYNTHETIC DATASET)

Reference:

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

[18] See https://www.cdc.gov/nchs/icd/icd10cm.htm

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 AND **117** COLUMNS

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	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 - $140,450	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	Prefer not to say
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	Yes
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	No
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	Yes
10	NO	45-49	1- 2 years	No	Certificate level, including trade	Ongoing/Permanent (other than senior executive)	Female	$183,300 - $261,450	No	No
11	NO	60-64	2 - 5 years	No	Bachelor Degree Level	Labour hire	Female	$183,300 - $261,450	No	Prefer not to say
12	NO	30-34	More than 20 years	No	Bachelor Degree Level	Ongoing/Permanent (other than senior executive)	Female	$183,300 - $261,450	No	No
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	$121,917 - $140,995	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	No
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	No
20	Prefer not to say	20-24	10 - 20 years	No	Advanced Diploma or Diploma level	Ongoing/Permanent (other than senior executive)		$151,763 - $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 68. SAMPLE OF SYNTHETIC NSW PMES DATASET

Reference:

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

DATASET 9 – SYNTHETIC NSW WORKFORCE PROFILE DATASET (SYNTHETIC DATASET)

Randomly generated dataset based on Public Service Commission[19] data, with fields including personal information (birth date, 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 AND **15** COLUMNS

300,000 UNIQUE SYNTHETIC EMPLOYEES (BASED ON GEN_CODE)

Gen_Code	DOB	Work and Live in Same Location Flag	Gender Code	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	Renumeneration_ 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	1717.609493

FIGURE 69. SAMPLE OF SYNTHETIC NSW WORKFORCE PROFILE DATASET

[19] See https://www.psc.nsw.gov.au/

Thanks

This paper was the culmination of more than three years' work by a Taskforce that 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, Internet of Things Alliance Australia, Data Synergies, Creator Tech, Objective, EY, Microsoft, Clayton Utz and several other companies.

Special thanks go to the contributors to the two Directed Ideation sessions:

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, Geof Heydon, Jakub Nabaglo, Wilko Henecka, Mariki Prozesky, Nathan Brewer, Richard Carlstein, Artem Kamnev, Huaifeng Zhang, Dominic Guinane, Stephen Katulka, Viki Ginoska, Aaron Butler, Dina Zebian, Fintan Guckian, Matthew Sohar, Sonya Sherman and Khimji Vaghjiani.

Thanks also go to the contributors to many, many "Safe" workshops over three years including:

Stephen Hardy, Peter Leonard, Chris Radbone, Geof Heydon, Sonya Sherman, Mathew Baldwin, Geoff Neideck, Frank Zeichner, Lyria Bennett Moses, Malcolm Crompton, Geoff Clarke, Kate Cummings, Ghislaine Entwisle, Ghazi Ahamat, Ben Hogan, Scott Nelson, Adrian Watson, Rachael Fraher, Alex Harrington, Andy West, Angelica Paul, Ashton Mills, Ben Hogan, Brian Thorne, Bridget Browne, Cassandra Gligora, Chris Mendes, Daniel Marlay, Dominic Guinane, Kelda McBain, Liz Bolzan, Luke Giles, Marilyn Chilvers, Matthew Roberts, Matthew McLean, Michael Wright, Mike Willett, Peter Hatzidimitriou, Rick Macourt, Robin van den Honert, Roulla Yiacoumi, Shona Watson, Suyash Dwivedi and Tiffany Roos.

Special thanks to Nick Rodwell, Peter Chiu, Shaun Murphy, Michael Ka, Alex Byrganov for their follow-up technical and background efforts and coordinating expertise, and to Jessica Kashro and Marc Portlock for their organising skills.

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

About ACS

The Australian Computer Society is the professional association for Australia's Information and Communications Technology sector. More than 45,000 ACS members work in business, education, government and the community.

ACS has a vision for Australia to be a world leader in technology talent, fostering innovation and creating new forms of value. We are firmly vested in the innovative creation and adoption of best of breed technology in Australia, and we strive to create the environment and provide the opportunities for members and partners to succeed.

ACS works to ensure ICT professionals are recognised as drivers of innovation in our society, relevant across all sectors, and to promote the formulation of effective policies on ICT and related matters.

Visit www.acs.org.au for more information.

ACS
International Tower One
Level 27, 100 Barangaroo Avenue
Sydney NSW 2000

P: 02 9299 3666
F: 02 9299 3997
E: info@acs.org.au
W: www.acs.org.au