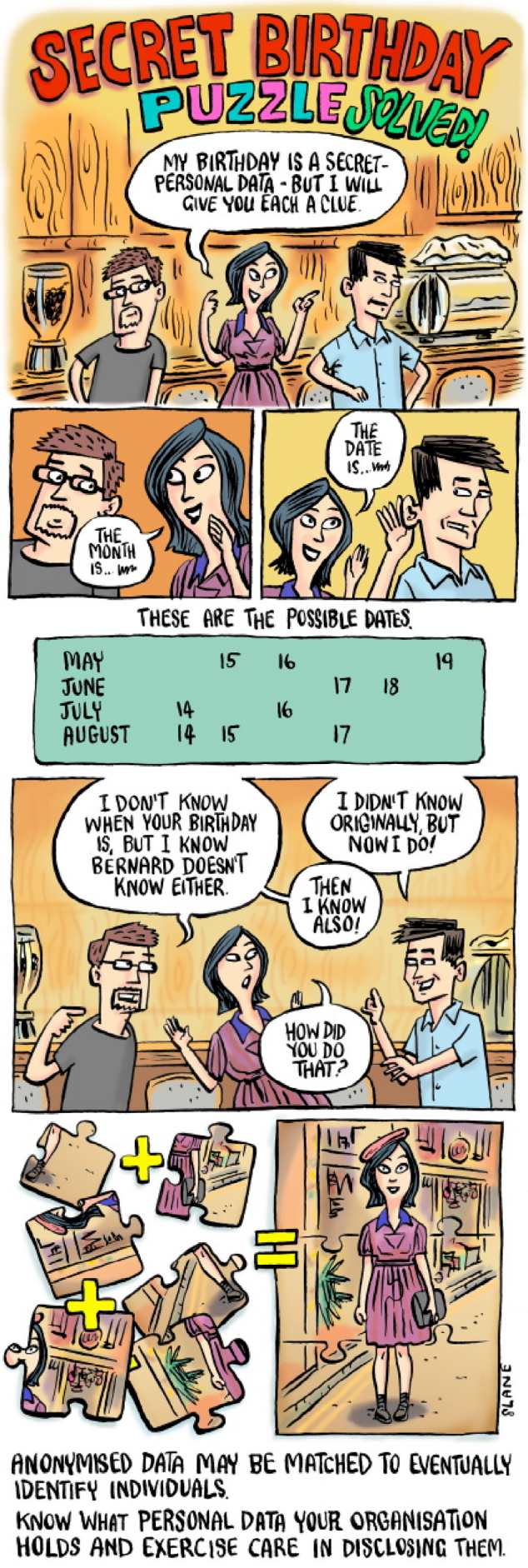

Anonymisation: Managing Personal Data Protection Risk

In April this year, an interesting math question on Cheryl's birthday put Singapore under international media spotlight after a Facebook posting about it went viral. Aside from igniting a debate on whether the question was too challenging for 14 year old students, it also highlighted a different set of conundrum from the perspective of personal data protection. The brain teaser superbly illustrated how personal information could be reconstructed from seemingly anonymised data.

Anonymisation generally refers to the process of removing identifying information such that the remaining data does not identify any particular individual. This is an important step to render the resultant data, which is no longer personal data, suitable for use in research and data mining. Such data analytics can bring greater value to different aspects of our lives, from improving transportation and healthcare services to enhancing public safety.

"There is often greater value in aggregating data instead of looking at specific data points," said Mr Zack Bana, Co-Founder and Data Protection Officer of Beacon Consulting, a management consultancy which conducts extensive research in areas such as customer satisfaction, brand perception, mystery shopping and employee engagement. Take, for example, a customer satisfaction survey which captures the income segment of individual respondents rather than their specific income. "With this information, I am able to draw certain conclusions, for instance, that individuals of a particular income bracket behave in a particular manner. That data is already anonymised, and we will not be singling out specific individuals."

"There is often greater value in aggregating data instead of looking at specific data points."

- Mr Zack Bana,
Co-Founder and Data Protection Officer of Beacon Consulting

Masking	Certain details of a piece of data are removed while preserving the essential look and feel of the data. ***Example:*** *Anonymising information on credit card numbers by replacing all but the last four digits by "X", or on NRIC numbers which may be represented by "S0XXXX45A" instead of the full set of original digits.*
Pseudonymisation	Information that can be used to identify a person is replaced with other randomly generated values from which the person's identity cannot be inferred. ***Example:*** *A person's name "Gerald Tan" may be represented by a random code "ABC869", and his NRIC numbers by "D346". The resulting dataset will still hold useful information on patterns and trends that is valuable to statistical research, while minimising the risk of individuals being identified.*
Aggregation	Values are displayed as a total figure instead of retaining the individual breakdown of numbers.
Replacement	An average figure is used to replace a value or a subset of values. ***Example:*** *For a subset of people with ages 15, 20 and 18, the value 17 may be used to blur the distinction if precision is not a requirement.*
Data suppression	A range is given instead of specific values.

Anonymising Data

There are numerous ways to go about anonymising personal data. In the birthday poser, Cheryl says her birthday is a secret, but gives Albert and Bernard two separate sets of clues as to when it might be. Albert is told it could be any one of four months. Bernard is told it could be any one of 10 days, of which only two occur uniquely.

This is an example of data reduction where some values are removed from a data set and, it is usually done because those values are not required.

Other common anonymisation techniques include masking, pseudonymisation, aggregation, replacement and data suppression.

An example of data suppression would be when Beacon Consulting is gathering data on age or income levels, respondents are often asked to indicate the range of values that they fall under, instead of providing their exact age or income. "This gives them the assurance that we will not be able to identify a particular respondent from the entire database," said Mr Andre Twang, the consultancy's Head of Research and Insights.

The consultancy is also mindful of the need for data anonymisation when it conducts focus group discussions in the course of a training programme. For example, there may be occasions where employees share statements and comments about their manager, work environment or organisation. "Sometimes, the examples quoted are very specific. If we are to use this information verbatim, it is likely that the organisation will know where this is coming from," said Mr Bana.

To ensure that it preserves anonymity in these scenarios, one of the things that Beacon Consulting does is to group an employee's inputs with similar comments made by other individuals so that it cannot be linked to a particular person.

Limitations and Challenges

There are often conflicting needs for anonymity and data integrity. Stripping data of too many identifiers may not preserve the usefulness of the data, or might deny potential uses for the data. Data anonymised for specific purposes might not be useful for others because its functionality is reduced.

For example, a retail organisation possesses a database of their customers' personal data (age, residential address, income, and occupation). From a marketing research perspective, these identifiers may yield information that is essential for profiling the customers and the development of personalised services. If the dataset were anonymised such that the ages were aggregated, incomes were shuffled, and the occupations removed, then the functionality would likely be lost as the anonymised database

would, while still providing general trends, not yield conclusions about customers' profiles, unlike the original database. The trade-off would then be the personalised services the organisation could have developed for their customers. It is, therefore, crucial to consider whether the anonymised data would still be suitable for its intended purposes.

There can also be limitations to anonymised data. For example, when using data suppression to group personal data into a certain range of values, there can be individual outliers who are instantly identifiable from the data. If a data set containing the ages of individuals has an outlier of age 89 while the other ages are below 50, the data point for the 89-year-old will stand out no matter how the ages are suppressed into ranges.

Another challenge is the risk of re-identification, brought about by the expanding pool of publicly available data that may be correlated to identify specific individuals.

In the case of Cheryl's birthday puzzle, Albert and Bernard are able to deduce that Cheryl's birthday falls on 16 July by putting together the two separate sets of clues that they have.

In a real-world example[1], graduate student Latanya Sweeney, who later went on to become a computer science professor, famously identified a Massachusetts governor by combining two data sets – the first from the Group Insurance Commission (GIC) which included zip codes, sex and birth dates of employees (with names, addresses and social security numbers removed) and the second from voter rolls which included names, zip codes, addresses, sex, and birth dates of voters in Cambridge, where the governor resided.

From GIC's database, only six people in Cambridge were born on the same day as the governor, half of them were men, and the governor was the only one who lived in the zip code provided by the voter rolls. The resultant data re-identified the governor and revealed information about his health and medical diagnosis.

What this example tells us is that while an organisation may consider a data set anonymised, it also has to consider the risk of re-identification if it intends to publish or disclose the data set to another organisation or make it publicly available.

Managing Re-Identification

To manage re-identification risks, organisations should consider if the entities receiving the anonymised data are likely to possess or have access to information that can inadvertently lead to re-identification. Personal knowledge is also an important factor in assessing re-identification risks, as the people who are close to an individual, such as an individual's friends or relatives, will possess unique personal knowledge about the individual. Although this personal knowledge will make it easier for an individual to be identified from an anonymous dataset by his or her friends than a stranger, it is unlikely to amount to high re-identification risks for the anonymised dataset.

Data should also be properly safeguarded from unintended recipients, whether they are within or outside the organisation.

Besides anonymisation, other practices that organisations can adopt to minimise the risks of re-identification include:

- Impose additional enforceable restrictions on the use and subsequent disclosure of the data
- Implement processes, including access restrictions, to govern proper use of the anonymised data
- Implement processes and measures for the destruction of data as soon as they no longer serve any business or legal purpose

Anonymising in a World of Big Data

Anonymisation remains a key tenet in personal data protection, safeguarding individual identities while allowing organisations to use data to gain valuable insights in more ways than would have been permitted under data protection regimes. As businesses embrace big data and use data analytics to do more sense-making and predictive analysis to extract insights to serve customers better and find new growth opportunities, anonymisation with robust re-identification assessment and risk management will be important to allow firms to optimally extract value from data and, at the same time, safeguard personal data.

To learn more about anonymisation, you can refer to Chapter 3 of the Advisory Guidelines on the Personal Data Protection Act for Selected Topics, issued by the Personal Data Protection Commission.

[1]Source: http://dataprivacylab.org/dataprivacy/projects/kanonymity/index.html