Office of the Victorian Information Commissioner

Disclosure of myki travel information

Investigation under section 8C(2)(e) of the *Privacy and Data Protection Act 2014* (Vic)

15 August 2019

Authorised by the Victorian Information Commissioner

Published by the Office of the Victorian Information Commissioner
PO Box 24274
Melbourne, Victoria, 3001

t: 1300 006 842
e: enquiries@ovic.vic.gov.au
w: ovic.vic.gov.au

Disclaimer
This publication may be of assistance to you, but the Office of the Victorian Information Commissioner and its employees do not guarantee that the publication is without flaw of any kind or is wholly appropriate for your particular purposes and therefore disclaims all liability for any error, loss or other consequence that may arise from you relying on any information in this publication.

Date of publication
15 August 2019

Report on disclosure of myki travel information

Good government policy and service planning relies on good evidence. We are fortunate that our public sector has access to such evidence in the form of high-quality datasets, many of which have been created or collated through public effort. Advances in analytics allow the sector to gain valuable insights from this data to deliver better services for the benefit of all Victorians. This can be a powerful tool for good and its responsible use is to be encouraged.

However, with such power comes responsibility. We need to ensure that the use of datasets does not weaken our human rights, including the right to privacy. While data-driven insights can bring great benefit, they can also put individuals at risk, particularly where datasets are made broadly accessible. The risks are often greater for those individuals that are already particularly vulnerable.

This report on OVIC's investigation into the release of myki data demonstrates that deficiencies in governance and risk management in relation to data can undermine the protection of privacy, even where the project is well-intentioned. The report also highlights that some of the assumptions made about data de-identification and release several years ago need to be revisited. Where a data set contains unit-level data about individuals, especially where it contains longitudinal unit-level data about behaviour, more recent research indicates such material may not be suitable for open release, even where extensive attempts have been made to de-identify it.

The recommendations in this report are aimed at ensuring that Victoria can continue to reap the benefits of data analysis while still respecting privacy.

This report has involved a significant investment of time and resources. I would like to thank the agencies that cooperated in the investigation for their assistance. I am releasing the report for its educative value, in the hope that it will help show a pathway to the responsible use of data insights to inform policy and service delivery decisions for the benefit of all Victorians.

Sven Bluemmel
Victorian Information Commissioner

Definitions

CPDP	Office of the Commissioner for Privacy and Data Protection
CSIRO	Commonwealth Scientific and Industrial Research Organisation
Deputy Commissioner	Privacy and Data Protection Deputy Commissioner, OVIC
DPC	Victorian Department of Premier and Cabinet
Information Commissioner	Victorian Information Commissioner, OVIC
IPP	Information Privacy Principle/s. Schedule 1, PDP Act.
OVIC	Office of the Victorian Information Commissioner
PDP Act	*Privacy and Data Protection Act 2014* (Vic)
PIA	Privacy Impact Assessment
PTV	Public Transport Victoria
The dataset	the myki dataset disclosed for use in the Datathon
The Datathon	Melbourne Datathon, held by Data Science Melbourne

Executive summary

1. In or around July 2018, Public Transport Victoria (**PTV**) released a dataset containing 1.8 billion historical records of public transport users' activity (the **dataset**) to a group known as Data Science Melbourne for use in the 'Melbourne Datathon' (the **Datathon**). The dataset contained the records of 'touch on' and 'touch off' activity of 15.1 million 'myki' cards used over a three-year period up to June 2018.

2. myki is a reusable electronic card used to pay for travel on metropolitan trains, trams and buses, myki-enabled v/line commuter trains, and regional buses in Victoria, Australia.

3. The Datathon is an annual event organised by Data Science Melbourne, at which participants compete to find innovative uses for a dataset. The Datathon commenced on 24 July 2018, running until 26 September 2018.

4. PTV stated that the dataset was disclosed to Data Science Melbourne in response to a request from the Department of Premier and Cabinet (**DPC**), which administers the Victorian Government open data platform through the 'DataVic' Access Policy and Guidelines. PTV stated that the purpose of disclosure was to support the Datathon.

5. Some steps were taken by PTV to de-identify the dataset before public release and to consider any associated privacy risks. PTV's conduct of a Privacy Impact Assessment (**PIA**) was premised on the assumption that the dataset had been successfully 'anonymised' by PTV, concluding that the dataset could therefore be safely released for use in the Datathon. This view, that the dataset had been de-identified, formed the basis for the governance of the released data.

6. During the Datathon, a participant raised concerns with a Victorian public sector representative that the dataset could be used to identify individuals. As a result, PTV were made aware of re-identification concerns and notified the Office of the Victorian Information Commissioner (**OVIC**) on 14 September 2018.

7. Separately, academics working at the University of Melbourne had located the dataset online and had been able to identify themselves, and persons known to them, in the dataset. On 20 September 2018, the academics notified OVIC of their findings and raised concerns regarding the release of the dataset, including the potential for numerous re-identification attacks on the dataset to be successful.

8. The Privacy and Data Protection Deputy Commissioner (the **Deputy Commissioner**) was concerned the publication of the dataset may present a risk to members of the Victorian community whose information could be re-identified, and that the release of the dataset might constitute a serious contravention of the *Privacy and Data Protection Act 2014* (Vic) (**PDP Act**) by PTV. As a result, the Deputy Commissioner decided she would initiate a formal investigation under section 8C(2)(e) of the PDP Act, to determine whether she should issue a compliance notice against PTV.

9. On 8 October 2018, the Deputy Commissioner formally wrote to PTV and DPC informing them of the investigation into the release of the dataset to Data Science Melbourne.

10. During the preliminary stages of OVIC's investigation, the Deputy Commissioner engaged data science experts from Data61, a division of the Commonwealth Scientific and Industrial Research Organisation (**CSIRO**), to further examine the dataset.

11. Data61's analysis was that the detailed nature of the information in the dataset created a high risk that some individuals may be re-identified by linking the dataset with other information sources.

12. Therefore, the Deputy Commissioner considered whether PTV breached Information Privacy Principle (**IPP**) 2.1 and 4.1 of Schedule 1 of the PDP Act in providing the dataset to Data Science Melbourne for the purposes of the Datathon.

13. The Deputy Commissioner found there were flaws in the process followed by PTV in de-identifying the dataset, assessing the risk of re-identification and deciding to provide the dataset for use in the Datathon.

Personal information in the dataset.

14. Information is 'personal information', and therefore subject to the PDP Act, where it is 'about' an individual whose identity is apparent, or can reasonably be ascertained, from the information.

15. The Deputy Commissioner's assessment is that the dataset contains information about individuals; namely, the location of people at specific times they started or concluded a public transport trip. The dataset also allows more information to be inferred about those people, such as their typical public transport movement patterns.

16. The Deputy Commissioner concluded the information contained in the dataset was personal information and must be handled in accordance with the IPPs in the PDP Act.

Breaches of the PDP Act

17. As PTV is required under the PDP Act to protect personal information in the dataset, it is the Deputy Commissioner's view that PTV breached IPP 2.1 by disclosing personal information for a purpose other than that for which it was collected. The Deputy Commissioner further considers that no exception to IPP 2.1 permitted the disclosure of the personal information in the dataset.

18. In disclosing the dataset to Data Science Melbourne in or around July 2018, the Deputy Commissioner found PTV contravened IPP 2.1 and therefore interfered with the privacy of the individuals whose personal information was in the dataset.

19. The Deputy Commissioner is also of the view that PTV breached IPP 4.1 in failing to take reasonable steps to protect the personal information contained in the dataset from disclosure. The steps taken by PTV in both considering Data Science Melbourne's request for the provision of myki data, and in preparing the dataset for release and use in the Datathon, were inadequate and not reasonable to protect the information contained in the dataset.

Outcomes for Public Transport Victoria, the Department of Premier and Cabinet and other public sector personal information custodians

20. There are several lessons arising from this matter; for PTV, for DPC, for the Victorian information regulator OVIC, and for other data custodians.

21. Principally, this matter demonstrates the challenges in identifying privacy risks in large, complex datasets and the need for the Victorian public sector, which possesses many large and sensitive data holdings, to have a high level of data literacy.

22. Secondly, appropriate processes and expertise should sit behind any decision to release de-identified personal information. PTV's decision-making processes were not clear or well documented and appeared to lack both the support of an effective enterprise risk management framework and suitable rigour in the application of a risk management process.

23. Throughout the process of developing and disclosing the myki dataset to the Datathon, and OVIC's investigation, both PTV and DPC displayed a lack of clarity about who was responsible for protecting the dataset and identifying and managing privacy risks.

24. This report makes recommendations to PTV and the Victorian public sector more generally. OVIC also considers it could have provided better regulatory guidance. These recommendations are outlined in paragraphs [184] to [203] of this report.

Decision to issue a compliance notice

25. The Deputy Commissioner considered PTV's submissions to this investigation, and all of the other material described in this report, before deciding PTV's breach was a serious contravention of the IPPs under section 78(1)(b)(i) of the PDP Act, and that a compliance notice should be issued. In reaching this view, the Deputy Commissioner considered factors including:

- the type of information in the dataset;
- the amount of information involved, and the number of people to whom it relates;
- the extent of harm to individuals and the likelihood of further harm that may result from the incident;
- the potential impact of the breach on public trust;
- PTV's response to the incident and its conduct during the investigation;
- PTV's willingness to implement the Deputy Commissioner's proposed recommendations;
- PTV's views on the definition of 'personal information' and related matters; and
- the fact that, to the best knowledge of the Deputy Commissioner, this was the only such incident involving PTV and PTV has not previously been subject to regulatory action from OVIC or its predecessors.

Decision to publish a report

26. The Information Commissioner considered a range of factors in considering whether the public interest requires the publishing of a report. These factors include:

- the need to provide transparency to the community about the issue, to allow the community to understand both the issue and the response taken by the public sector;
- the educative value of publishing an investigative report for PTV, DPC, OVIC and other data custodians;
- the potential for a public report to lead to better decisions on open data; and
- a consideration the dataset vulnerability was likely to come to wide public attention at some point, and that it was preferable that it do so in the context of a regulatory investigation, and a compliance notice requiring remediation action.

27. On balance, the Information Commissioner decided it was in the public interest to publish a report under section 111(3) of the PDP Act.

Background

28. The Melbourne Datathon is an annual event at which members of the Victorian data science community compete to find innovative uses for a dataset. The Datathon is the largest event of its nature in Australia. For the 2018 Datathon, the dataset being examined was a record of public transport trips recorded on the myki[1] ticketing system over a three-year period, from 1 July 2015 to 30 June 2018. The dataset was provided to the Datathon by PTV, Victoria's principal public transport agency.[2]

29. The dataset recorded 'touch on' and 'touch off'[3] events on the myki public transport ticketing system. The data was released as a linked unit level dataset, consisting of records of individual transactions (trips) linked through a numerical identifier assigned to each card. The numerical identifier allowed a travel history for that card over the three-year time period of the dataset to be constructed.

30. During the Datathon, a participant advised a Victorian public sector representative that it may be possible to identify individuals within the dataset and determine their public transport movements. Separately, academics from the University of Melbourne downloaded the dataset from the Datathon's website (where it was published for Datathon participants) and used the data to identify their own travel movements and the travel movements of other people known to them. The academics notified OVIC of their concern. OVIC contacted PTV and DPC who were both involved with the Datathon. By this time, PTV and DPC had already formed a response team to consider the concern raised by the Datathon participant. After the Datathon ended, the dataset was taken off the website.

31. The Deputy Commissioner was concerned PTV's release of the dataset may have exposed the personal information of people who had travelled on the myki network. Based on the analysis completed by the University of Melbourne academics, it appeared travel records in the dataset could be linked to individuals' identities in some circumstances. This could allow a malicious third party with access to the data to determine another individual's history of public transport journeys.

32. OVIC considered members of the Victorian community would expect information about their travel movements to be afforded a high degree of protection. OVIC also considered potential scenarios in which misuse of this information could lead to adverse consequences for individuals if their information was revealed. The amount of data also suggested this was a serious issue: the data recorded the travel movements of most people who had used the myki public transport system during the three-year period. This amounted to millions of Victorians.

---

[1] myki is a contactless smartcard ticketing system operated by PTV, used for electronic payment of public transport fares in Melbourne and parts of regional Victoria. Public use of myki in Melbourne commenced in December 2009 for Melbourne metropolitan train services. From December 2012, myki was the only valid ticket for Melbourne public transport. PTV owns the trademark 'myki'. The lack of capitalisation used in the PTV trademark is adopted in this report.

[2] From 1 July 2019, PTV's functions were moved into the Victorian Department of Transport. For convenience, 'PTV' refers to whichever agency carries out functions under s 79AE(1)(k) of the *Transport Integration Act 2010* (Vic) at the relevant time.

[3] 'Touching on' and 'touching off' events refer to the use of the myki electronic ticket card. The public transport user 'touches' the card on an electronic card reader when starting or ending a public transport journey. This action both allows payment to be made for the journey, and also allows access to and from public transport sites, often through swing gates. 'Touch[ing] on' and 'touch[ing] off' is also referred to as 'tap[ping] on' or 'tap[ping] off' the myki card.

Due to these factors, the Deputy Commissioner initiated an investigation to consider whether PTV had contravened the PDP Act in releasing the information.

33. This section provides background to the data release, including key events leading to the release of the myki dataset, concerns being raised with OVIC and PTV about re-identification risks for the dataset, and steps taken by PTV and DPC after those risks were drawn to their attention.

myki data held by Public Transport Victoria

34. myki is a reusable electronic card used to pay for travel on metropolitan trains, trams and buses, V/Line commuter trains and myki enabled regional buses. The myki card registers touch-on and touch-off data to record payment for transport. A myki must be touched on for it to be valid for a journey or entry to a compulsory ticket area. A myki may need to be touched off depending on mode of transport, relevant fare, and ticketing conditions.

35. Although held and used by commuters, myki smartcards are legally the property of the head of Transport for Victoria. PTV and its authorised representatives may inspect, suspend, or take possession of a myki smartcard, or require its return, at any time.

36. At the request of an individual who uses a myki smartcard, PTV or Transport for Victoria will register a myki to that individual. An individual can register up to eight myki cards on an individual account. This permits one person to manage myki cards for family or friends. In certain circumstances, a myki may be shared between more than one person.[4]

Public Transport Victoria, the Department of Premier and Cabinet, and Data Science Melbourne

37. PTV was a Victorian statutory authority which, until 1 July 2019, managed Victoria's public transport network and the myki ticketing system. From 1 July 2019, PTV's functions were moved into the Department of Transport. For convenience, whenever this report refers to 'PTV' it is referring to the statutory agency up to 1 July 2019, and the Department that carried on the relevant functions of that agency after that date.

38. One of PTV's functions, under section 79AE(1)(k) of the *Transport Integration Act 2010* (Vic), is to provide and operate, or facilitate the provision and operation of, ticketing systems used for the public transport system and manage ongoing improvements in the ticketing systems for the public transport system.

39. DPC is one of eight Victorian public sector departments. It supports the Premier of Victoria and leads and coordinates the activities of the Victorian public service.[5] It is also responsible for a number of policy areas and programs. One of the programs administered by DPC is DataVic, an online repository of Victorian public sector open datasets. DPC is responsible for promoting the DataVic Access Policy, which promotes the sharing and release of Victorian public sector

[4] A myki may be shared, in summary, if it the myki was not issued with a free travel pass, or during a period that it has an active myki pass (a pre-paid ticket which permits unlimited travel in certain areas for a specified time): see PTV, 'Victorian Fares and Ticketing Manual', 1 July 2019, pp 6-7. A myki may also be purchased but not linked to an individual account – but in this case it may not have funds added via an online process. Concessional myki cards must be linked to individual accounts.

[5] https://vic.gov.au/department-premier-and-cabinet/.

datasets.[6] DPC was represented on the Datathon judging panel and provided sponsorship funding to Data Science Melbourne in support of the 2018 Datathon.[7]

40. Data Science Melbourne is a Meetup[8] group organised by and for people with an interest in data science.[9] Data Science Melbourne organises the Datathon, an annual event in which participants compete to find uses for a dataset provided to them by the event organisers and sponsors.[10] The Datathon is sponsored by a number of commercial enterprises, educational institutions, and, in the case of the 2018 Datathon, the Victorian Government.[11]

Use of myki data at the Melbourne Datathon

41. In 2015, Data Science Melbourne approached DPC to request the release of a dataset in support of the 2016 Datathon. Data Science Melbourne identified myki data specifically as a dataset of interest. DPC contacted PTV on behalf of Data Science Melbourne to request access to the myki data, but the request was declined because of concerns about ownership of the data.

42. In or around December 2017, DPC again approached PTV with a request to support the Melbourne Datathon by providing data about public transport trips made on the myki ticketing system.[12] A meeting was held between DPC, PTV, and Data Science Melbourne to discuss this proposal on 14 December 2017. At the meeting, DPC introduced PTV to Data Science Melbourne. Based on that meeting, PTV understood Data Science Melbourne was organising the Datathon on behalf of DPC. In fact, DPC had a more limited involvement.[13]

43. Following the meeting, PTV considered Data Science Melbourne's request. On or about 17 January 2018, PTV completed a PIA to assess whether the dataset could be used in the Datathon. The PIA concluded the dataset could be modified to allow use in the Datathon without disclosing personal information. The PIA did not describe exactly what data would be released, other than to say it would be anonymised myki data. The PIA was approved by the PTV 'owner' of the myki dataset, and the PTV chief information officer. The PIA was the only authorising decision or documentation for PTV's decision to release the data.[14]

44. On or about 18 January 2018, PTV informed DPC they had completed a PIA, which provided 'the OK to release the myki data' to Data Science Melbourne.

45. Between January and June 2018, PTV staff and Data Science Melbourne discussed the details of exactly what data would be included in the dataset. During this time a number of smaller sample datasets were provided to Data Science Melbourne to indicate the data that was available, and to confirm the data would be fit for purpose. These initially appeared to focus on short sample periods, or samples of geographic areas. Over time, the scope and quantity of data that would be provided was clarified and expanded. A fuller dataset would be more useful for the Datathon. Eventually, PTV and Data Science Melbourne agreed a three-year window of

[6] 'Data Vic Access Policy', available online at https://data.vic.gov.au/datavic-access-policy.
[7] 'Victorian Common Funding Agreement Ref D18/107723', 14 June 2018.
[8] Meetup is an online platform used to organise groups that host in-person events for people with similar interests.
[9] https://www.meetup.com/en-AU/Data-Science-Melbourne/.
[10] http://www.datasciencemelbourne.com/datathon/.
[11] http://www.datasciencemelbourne.com/datathon/#sponsors.
[12] 'Re: PTV notification following a privacy query regarding myki data set in Melbourne Datathon 2018', 14 September 2018 ('**PTV notification**'), 'Background'.
[13] See discussion below, from paragraph [174] onwards.
[14] See discussion below, from paragraph [138] onwards.

all public transport trips, rather than a smaller sample, would be provided to Data Science Melbourne.

46. On 27 April 2018, Data Science Melbourne wrote to DPC to seek sponsorship for Datathon prizes. On 26 June 2018, DPC and Data Science Melbourne signed a grant agreement under which DPC provided grant funding to offer prizes to Datathon participants. This agreement did not touch on the use of the myki dataset provided by PTV. This was the only written agreement between Data Science Melbourne and the Victorian public sector relating to the Datathon.

47. On 16 May 2018, Data Science Melbourne requested confirmation by email with DPC and PTV that there were no restrictions on Data Science Melbourne or Datathon participants that would limit their use of, or prevent them from keeping, the dataset. PTV provided this confirmation on 23 May 2018. The absence of restrictions on the data was highlighted to participants on the Datathon website: 'No NDA [non-disclosure agreement] to sign this year – you can do what you like with the data.'[15]

48. On 12 July 2018, PTV provided the final version of the dataset to Data Science Melbourne.

49. The dataset included myki 'touch on' and 'touch off' data and look-up tables for stop location and card type. It contained 1.8 billion historical myki records of 'touch on' and 'touch off' activity associated with 15.1 million myki cards, from the three-year period up to June 2018. It also recorded some information about individual myki cards – most notably, the 'concession type' of cards issued to provide discounted or free travel to certain groups (for example, children, seniors, refugees, police and politicians).

50. The Datathon commenced on 24 July 2018 running until 26 September 2018.

Discovery of re-identification risk

51. Data Science Melbourne held a briefing for Datathon participants on 13 September 2018 to provide them with information about the dataset and the Victorian public sector's open data program. At the briefing, a DPC representative (appearing because PTV's representative was unable to attend) gave a presentation to the Datathon participants about the dataset. After the presentation, a participant approached the DPC representative to raise concerns about the dataset. These included that:

    - the dataset identified myki card types. The participant raised a query about the potential identification of state and federal police and politicians in the data. The participant noted there were very few politician concession cards, which would make linking these cards to individuals easier; and
    - the participant was able to identify the travel movements of a friend in the dataset, based on knowledge of some trips that friend had taken. By identifying those known trips, the participant could identify all other trips that used the same card.

52. On 20 September 2018, Dr Chris Culnane, an academic at the University of Melbourne, contacted OVIC to raise concerns about the release of the myki dataset. Dr Culnane attached a paper outlining how he and two co-authors (Dr Benjamin Rubinstein and Dr Vanessa Teague) had located the dataset online, and had been able to identify themselves, and third parties, in the dataset. They downloaded the data from an open Amazon Web Services 'S3' "bucket" linked to the Datathon's public facing website. They stated they had re-identified themselves

[15] 'How it Works', Melbourne Datathon website. http://www.datasciencemelbourne.com/datathon/.

and others by matching information in the dataset with information about known public transport journeys.[16] The paper outlined that its authors had:

- re-identified their own travel records. Two of the academics had registered their myki cards. This meant they were able to access historical, 'to the second' trip data for the previous six months through PTV's website. They matched this with information in the dataset. To confirm they had found their own records, they cross-checked the registration dates of the cards with the first recorded trips.
- Successfully re-identified a 'co-traveller' who had travelled with one of the academics on a single occasion. They did so by identifying everyone who had touched on or off the relevant tram at about the same time as the academic, and then used their knowledge of the person's general work and home location to narrow down potential candidates by looking at their travel patterns. To confirm they had re-identified the correct person, they cross-matched the record with some further information obtained from the person (the expiry date of the myki card). The academics said, 'this type of re-identification is particularly concerning, since it allows an individual to leverage the ease of re-identifying themselves to re-identify others, and from potentially only a single co-travel event'.
- Noted there may be additional re-identification attacks possible against particularly vulnerable groups, because the dataset discloses card types. These include parliamentarians, police officers, and children. They noted some card types have very few individual myki cards issued against them (for example, the Federal Parliamentarian card type has only seven registered cards).

Response to the incident by Public Transport Victoria and the Department of Premier and Cabinet

53. After being contacted by the Datathon participant, DPC and PTV formed a response team to consider the participant's claim. The team investigated the participant's allegations but concluded there was not a significant risk arising from the dataset's release. PTV and DPC's response team found 'the dataset had been anonymised by PTV prior to release, with a Privacy Impact Assessment completed', and 'it was not possible using the myki dataset alone to positively identify specific travellers and their prior travel movements. Supplementary information is required before positive identification can be made.'

54. DPC notified the Australian Federal Police and Victoria Police, because the participant's claim was about risks to police officers (who, in Victoria, are given free travel, and issued a special category of concession card). Victoria Police conducted a risk assessment and reported there was a 'low/minimal' risk to the safety of police and parliamentarians as a result of the data release.

55. DPC initiated a review of the policy and procedures underpinning the operation of the Melbourne Datathon 2018 to identify how these events can better meet community and stakeholder expectations.

The data release and Victoria's Open Data Program

56. In 2012, the Victorian Government published the *DataVic Access Policy*, which provides a 'plan for enabling public access to government data'.[17] It includes five principles to support

[16] Chris Culnane, Ben Rubinstein, Vanessa Teague, 'Myki Re-Identification' ('**Re-identification paper**'), 20 September 2018. OVIC was provided a draft of this report.
[17] DPC, 'DataVic access policy', <data.vic.gov.au/datavic-access-policy>.

appropriate release of data. The policy says more public access to government data will stimulate economic activity and drive innovation, increase productivity, improve research outcomes, and improve the efficiency and effectiveness of government. Principle 1 states:

Government data will be made available unless access is restricted for reasons of privacy, public safety, security and law enforcement, public health, and compliance with the law.

57. As noted above, the DataVic Access Policy is administered through DPC. Its principles are supported through more detailed guidelines in the form of the DataVic access policy guidelines.[18] These guidelines provide, among other things, guidance about when data derived from personal information should be released, and how to avoid the re-identification of data.

58. During the investigation, PTV and DPC indicated the release of the dataset was in support of the DataVic open data policy, and that the dataset as released had been intended to be open data – explaining the lack of limitations on use and reuse. However, PTV's PIA did not appear to envisage the dataset being released as open data, and other contemporaneous documents provided mixed accounts of how the data was intended to be released or used. Furthermore, the data release, and the manner in which the data was de-identified, did not accord with the DataVic access policy guidelines. This issue is discussed in more detail later in this report.

[18] Available online at <data.vic.gov.au/datavic-access-policy-guidelines>.

Investigation by the Office of the Victorian Information Commissioner

59. OVIC considered it was necessary to better understand the extent of the privacy risk caused by the disclosure of the dataset and whether this indicated a breach of the IPPs in Schedule 1 of the PDP Act. The Deputy Commissioner commenced an investigation on 8 October 2018.

Decision to investigate

60. OVIC became aware of the issue with the myki dataset in two ways. First, PTV notified OVIC of the concerns raised by the Datathon participant. Second, and independently, academics from the University of Melbourne contacted OVIC.

61. The Deputy Commissioner was concerned the publication of the dataset may present a risk to members of the Victorian community whose information could be re-identified, and that the release of the dataset might constitute a serious contravention of the PDP Act by PTV.

62. Under section 8C(2)(e) of the PDP Act, the Information Commissioner or Deputy Commissioner can issue a compliance notice and carry out investigations for the purpose of deciding whether to issue a compliance notice.

63. Under section 78(1) of the PDP Act, a compliance notice may be served on an organisation if it appears that:

- the organisation has done an act or engaged in a practice in contravention of an IPP; and
- the act or practice –
    - constitutes a serious or flagrant contravention; or
    - is of a kind that has been done or engaged in by the organisation on at least 5 separate occasions within the previous 2 years.

64. Under Section 78(2) of the PDP Act a compliance notice requires an organisation to take specified action, within a specified period, for the purpose of ensuring compliance with the IPPs.

65. An investigation may also lead to the publication of a report and recommendations under section 111 of the PDP Act. Section 111 permits the Information Commissioner to publish a report where the Information Commissioner considers it is in the public interest to do so. The Information Commissioner may report on any act or practice the Information Commissioner considers to be an interference with privacy, or report about any matter generally relating to the Information Commissioner's function under the PDP Act.

66. On 8 October 2018, the Deputy Commissioner wrote to PTV to advise she intended to investigate the de-identification and disclosure of the myki dataset. The Deputy Commissioner considered it was appropriate to investigate this incident under sections 8B(1)(a), 8C(2)(e) and 78 of the PDP Act, having regard to the likelihood that a breach had occurred, the potential severity of the breach, and the response to the incident by PTV.

67. The Deputy Commissioner did not recommend PTV notify the community of the possible data breach. This was for two reasons. At the beginning of the investigation it was not clear to what extent any individuals were at risk because of the incident. Further, the re-identification method used by the University of Melbourne academics relied on linking information available

to myki travellers via their online PTV account, which provides travel records for the previous six months. The Deputy Commissioner considered it would be too great a risk to raise awareness of that possible re-identification method while the current six-month window overlapped with the date range of the published dataset.

Scope of investigation and issues to be considered

68. Section 20 of the PDP Act states an organisation must not do an act, or engage in a practice, that contravenes an IPP. PTV is an 'organisation' for the purpose of Part 3 (Information Privacy) of the PDP Act, as it is a body established for a public purpose by or under an Act.[19] The Deputy Commissioner advised PTV she considered the following IPPs were relevant to this investigation and would consider whether they had been contravened:

- IPP 2 (use and disclosure), which prohibits an organisation from using or disclosing personal information for a purpose other than that for which it was collected, unless an exception applies; and
- IPP 4.1 (data security), which requires an organisation to take reasonable steps to protect the personal information it holds.

69. OVIC's investigation considered whether PTV contravened the above IPPs. In addition, the investigation considered other matters relevant to a decision about whether to issue a compliance notice, and on what terms. This involved consideration of:

- the extent to which there was a risk that personal information of members of the Victorian community was exposed as a result of the incident;
- if there was a breach, the adequacy of steps that PTV (or other Victorian public sector agencies) had taken in response to the incident; and
- if there was a breach, what further steps, if any, should be taken by PTV or the Victorian public sector in response to the incident.

Information considered

70. Information was gathered by OVIC through meetings held between OVIC and PTV, DPC, the University of Melbourne, and Data Science Melbourne. OVIC made requests for documents and written responses to questions from PTV and DPC. The Deputy Commissioner also commissioned quantitative and qualitative analysis of the re-identification risk of the released dataset. This was conducted by Data61, a division of CSIRO.

71. PTV, DPC and Data Science Melbourne all cooperated fully with the Deputy Commissioner's investigation. PTV and DPC substantially assisted OVIC by responding to all questions and requests for documentation in a timely and comprehensive way. Their response to the investigation indicated an openness and willingness to respond constructively to the Deputy Commissioner's concerns.

72. In reaching and maintaining the views outlined in this report, the Deputy Commissioner considered the following material:

- written submissions from PTV and DPC;
- information gathered in meeting with representatives from DPC, PTV, and Data Science

[19] PDP Act s 13(1)(f).

Melbourne;

- correspondence between PTV, DPC and Data Science Melbourne in the lead-up to the Datathon;
- the contract between DPC and Data Science Melbourne for Datathon sponsorship;
- the PIA conducted by PTV for the Datathon;
- PTV privacy policies;
- documentation recording PTV's data governance program;
- documentation relating to DPC and PTV's incident and post-incident response;
- a re-identification risk assessment report conducted by Data61 on behalf of OVIC;
- myki re-identification report completed by academics at the University of Melbourne; and
- information obtained from the websites of PTV, DPC, and Data Science Melbourne.

The data release and Information Privacy Principle 2

73. IPP 2 (use and disclosure) prohibits an organisation from using or disclosing personal information for a purpose other than that for which it was collected, unless an exception applies.

74. To consider whether IPP 2 was contravened, OVIC firstly questioned whether the dataset disclosed to Data Science Melbourne contained personal information. The second question considered was whether the use and disclosure of the dataset was permitted by IPP 2, in light of the purpose for which the data was collected, and the circumstances in which it was used or disclosed.

Did the dataset contain personal information?

Personal information

75. Personal information is defined in section 3 of the PDP Act to mean:

> *information or an opinion (including information or an opinion forming part of a database), that is recorded in any form and whether true or not, about an individual whose identity is apparent, or can reasonably be ascertained, from the information or opinion.*

76. Under this definition, there are two steps in determining whether information or an opinion is personal information. The first step is to ask whether the information or opinion is 'about' an individual. If it is, the second step is to ask whether the identity of that individual 'is apparent or can reasonably be ascertained, from the information or opinion.'[20]

77. An important qualification in the definition of personal information is that a person's identity must be apparent or reasonably ascertainable 'from the information or opinion'. However, this does not mean the information in question is all that can be considered in deciding whether the information is 'personal information'. It is clear some extraneous material or information may be considered.[21] This has long been the approach that privacy regulators have taken to this definition. The 1983 report of the Law Reform Commission, which provided the definition of personal information that remains in Victoria to this day, foresaw the possibility of combining information to identify an individual:

> *[i]f the information can easily be combined with other known information, so that the person's identity becomes apparent, the information should be regarded as personal information. Information should be regarded as 'personal information' if it is information about a natural person from which, or by use of which, the person can be identified.*[22]

---

[20] *Telstra Corporation Limited and Privacy Commissioner* [2015] AATA 991 [97], affirmed in *Privacy Commissioner v Telstra Corporation Limited* [2017] FCAFC 4. Although these cases were decided with respect to the *Privacy Act 1988* (Cth), the definition of 'personal information' in the *Privacy Act 1988* (Cth) at the time was, in all material respects, the same as the definition in the PDP Act. The similarity is deliberate, in the interests of supporting a nationally consistent approach to the protection of information privacy: see *Explanatory Memorandum to the Privacy and Data Protection Bill*, page 3. For these reasons, these Commonwealth cases can be relied upon in interpreting the definition of 'personal information' in the PDP Act.

[21] As set out in *WL v La Trobe University* [2005] VCAT 2592, 45.

[22] Report No. 22, AGPS Canberra, 1983, Vol 2. The Law Reform Commission's Report preceded the *Privacy Act 1988* (Cth) and provided the definition of personal information adopted in that legislation. As noted in a footnote above, the definition of 'personal information' in the Victorian PDP Act was adopted from the *Privacy Act 1988* in the interests of consistency.

78. The definition of personal information in the PDP Act is deliberately broad, as it defines the limits of the PDP Act's application.[23] The intention behind adopting a broad definition is that wherever possible, the PDP Act will protect personal information subject to the operation of the IPPs and any exemptions.[24] The answer to the question 'what is personal information?' is an 'evaluative conclusion, depending upon the facts of any individual case'[25]. To determine whether a piece of information is 'personal information', it must be considered in context and on a case-by-case basis.

79. A dataset containing personal information may be modified to make it more difficult (or less likely) for an individual's identity to be ascertained from the information. Depending on the effectiveness of the method used to de-identify the dataset, and the context in which the information is subsequently used and held, the result may be that no individual's identity is 'apparent' or can 'reasonably be ascertained' from the information. If this is the case, the information no longer meets the definition of personal information as it is considered de-identified and no longer subject to the IPPs.[26]

Description of the dataset

80. The myki dataset contains a record of 'touch on' and 'touch off' events recorded by the myki system between 1 July 2015 and 30 June 2018, amounting to approximately 1.8 billion events across 15 million distinct myki cards. Each event record comprises multiple data points:

- **date and time** - the date and time of the 'touch on' or 'touch off' event, to an accuracy of one second;
- **location information** — this data varies depending on the mode of transport, but includes information such as vehicle identifiers, route numbers and stop numbers;
- **card identifier** — a unique number assigned to each myki card when preparing the dataset for release. The card identifier permits all 'touch on' or 'touch off' records within the dataset relating to a particular card to be linked; and
- **card type** — a descriptor of the type of card used. There are approximately 70 myki card types, including card types specific to Victoria Police, Federal Police, State and Federal Parliamentarians, asylum seekers, veterans and pensioners.

81. The card identifier field cannot be directly linked to the myki card number printed on the face of the card. Rather, it is a derived number created from an internal card identification number, used by PTV internal systems only. The card identifier used in the dataset was created by applying an algorithm to the internal card identification number to replace it with a different number. The new number was intended to be meaningless.[27]

82. The dataset was not a complete record of all myki trips, as it does not contain records for certain older myki cards with card types that do not match the card type categories used when

[23] *Jurecek v Director, Transport Safety Victoria* [2016] VSC 285 [78].
[24] *Jurecek v Director, Transport Safety Victoria* [2016] VSC 285 [78].
[25] *Privacy Commissioner v Telstra Corporation Limited* [2017] FCAFC 4 [63].
[26] This generally accords with the definition of 'de-identified' in s 3 of the PDP Act. However, it should be noted that the only place in the PDP Act that this defined term is used is in IPP 4.2, with respect to the obligation of organisations to destroy or de-identify information that is no longer required. As such, where this report refers to 'de-identify' or 'de-identified', it is not referring to the definition in s 3 of the PDP Act, but rather, to a treatment of information that results in the information no longer meeting the definition of 'personal information'.
[27] The method used to transform the card identifier is discussed further detail below at [159] to [162].

extracting the data. It appears these were omitted from the dataset disclosed to Data Science Melbourne due to an oversight, rather than as an intentional de-identification measure. In any event, the dataset includes all myki transactions for a large majority of myki cards used during the covered time period.

Submissions of Public Transport Victoria

83. It is clear that before releasing the dataset PTV had considered this issue and decided the dataset did not contain personal information. PTV maintains the dataset does not contain personal information. In November 2018, PTV submitted to OVIC that:

> *PTV does not consider the data extract is personal information as defined in the [PDP Act]. PTV's view is that there has been no breach or contravention of the Information Privacy Principles (IPPs) as result disclosing the data extract to the Datathon. This is based on our interpretation of the definition of personal information which [is] key to the establishment of the sensitivity of the data and therefore its impact if a breach did occur. The data extract disclosed for the Datathon contained no personal identifiable information. The ability to identify an individual [rests] with the relationship between the card number and the myki account for that individual. The data extract disclosed to the Datathon substituted each card number with randomly generated card numbers which anonymised the individuals. This was undertaken by PTV prior to the disclosure. As it is not possible to identify individuals through the card numbers disclosed to the [Datathon] this significantly lowers the sensitivity of the data disclosed.*

84. PTV provided several additional written submissions about whether the information was personal information. The Deputy Commissioner carefully considered each submission made by PTV. The above extract provides a good overview of PTV's position, and it is not necessary to set out PTV's submissions in full.

Technical analysis of the dataset

85. An analysis of the re-identification risk of the dataset was completed for OVIC by Data61. Data61 was engaged to analyse and describe the dataset, and provide an expert opinion about re-identification risks to the dataset. Data61's opinion is that the overall risk of re-identification for the dataset is 'extremely high'.

86. Pertinent points from Data61's analysis as to the question of whether the dataset contains personal information include that:

- the 'overall re-identification risk of the myki dataset is extremely high on reasonable knowable background information. It is only the uncertainty around the ability to know background information that may reduce overall risk from extremely high to high';[28] and
- the Dataset exhibited a high level of uniqueness. For example:

  > *[c]ombining two events by a card, on a typical day (we used 8 Feb 2017) based on time (to the second) and stop location (not even considering whether the event is a scan on or off), over 66% of scan events are unique. When time is generalised to 10 minutes, that risk drops, but still 5.5% of card scan pairs can be uniquely identified and many cards are in small groups that would allow someone to make an educated guess. When two scans by a card are known by time to 10 minutes and stop location, 61.9% of those pair of scans are unique. This illustrates the high risk when events are combined.*[29]

[28] Data61, 'Re-identification Risks Assessment for the Office of the Victorian Information Commissioner on the Public Transport Victorian myki Dataset provided to Melbourne Datathon 2018'.
[29] Ibid.

87. The report identified two hypothetical and one real re-identification scenarios. For example:

- a nosy co-worker or estranged spouse knows a persons' past travel movements. The report considered a situation where two 'tap on' [and] 'tap off' events in a single day were known, in terms of stop location, and time of tap on/tap off to the nearest ten minutes. In this situation, 61% of combinations were unique. Trips involving the five busiest Melbourne CBD stations had a far lower but still significant proportion of unique events (9.8%). However, as more events are known, the proportion of unique combinations quickly increases;
- a scenario involving a family holiday. The Data61 report authors were able to identify their own myki cards using only the fact they had travelled to the nearest public transport stop to five well known tourist attractions within an 11 day window; and
- identifying a card based on the first occasion on which it is used. The report authors used this scenario on the basis that 'knowledge of using a card for the first time could be easy to know through casual conversation, by observation, or by knowing someone was recently arrived in Melbourne.' The report noted 36% of first scans are unique, when the location of the scan and time at which the scan occurred to the nearest ten minutes is known.

88. The report by Data61 also stated the authors' opinions about the resources needed to re-identify the dataset and the possible motivation of people to do so. The authors' considered:

Resources needed for re-identification

The size of the myki dataset is large and unable to reside in memory on a typical desktop computer. This explains the splitting of the dataset that occurred at the Datathon. The analysis in this report has been conducted on a 40 core machine with 512GB of RAM. However these requirements have only been necessary as the analysis is looking for the possibility of re-identification across all people in the dataset (all myki cards) for a number of combinations of attributes. As is seen in some of the scenarios and most of the analysis, reducing the dataset to smaller time periods, particularly those a party may be interested in, and filtering based on stop locations would reduce the need for high powered computers.

The dataset could be easily split into time periods that are much shorter and so become well within the scope of a typical desktop or laptop computer. In this case, simple attacks searching for particular individuals in particular subcategories may be possible by a person with skills in Microsoft Excel, a good understanding of the dataset attributes, and a bit of spare time. For a more complex re-identification, the required skills and techniques are available (though rarer) in the work force. An experienced data analyst would have no trouble handling this dataset and conducting re-identification. Scenario 2 analysis was all conducted on a command line in a Linux distribution with common commands such as grep, awk, sort, uniq and sed.

Motivation of a Party and Likelihood of Access to Resources for Re-identification

This dataset is a highly attractive and easy to understand dataset. The ability to understand the dataset and to match background knowledge to a recorded trip is one that could be easily envisaged by someone with moderate computer skills. The attractiveness of learning an individual's myki card and their entire trip history over 3 years is high and anticipated to have broad appeal. We note that OVIC have already had interested parties indicate self re-identification.[30]

89. Analysis of the dataset was also conducted by Dr Chris Culnane, Dr Benjamin Rubinstein and Dr Vanessa Teague of the University of Melbourne. The academics provided a report to OVIC

[30] Ibid.

summarising their concerns. Particularly relevant points to the question of whether the dataset contains personal information are extracted above at paragraph [52].

90. Neither the Data61 report nor the University of Melbourne report provide an answer to the question of whether the dataset contains personal information. They are not designed to do so. The reports analyse and describe the dataset and provide the authors' expert opinions about re-identification risks to the dataset. To decide whether the information is 'personal information', it is also necessary to consider factors such as the context in which the dataset was released, and relevant legal authorities.

Findings

91. To determine whether the dataset contains personal information, it is necessary to consider two questions. First, does the dataset contain information about any individual? Second, is the identity of any of those individuals apparent, or can it be reasonably ascertained, from the information in the dataset?

Does the dataset contain information 'about' an individual?

92. Information will be 'about' an individual when it is 'on the subject of' or 'concerning' the individual.[31]

93. PTV's view is that the dataset is information about the service it provides 'to' individuals, and, regardless of the dataset containing 'touch on' and 'touch off' records, the information is not 'about' these individuals. While this information is about a card, the data also represents information which reveals the movement patterns of those cards across the public transport system during the relevant period. In most cases, the movement of a myki card will match the movement of the person who owns the myki card. Each 'touch on' or 'touch off' event in the dataset reveals the location of an individual at a particular time, and the fact they were starting or ending a public transport trip. These movements can be connected to create a more detailed picture of that individual's movements. This is information 'about' that individual, as well as being information about PTV's service.[32]

94. PTV also submitted that a myki card may be shared by multiple people.[33] Where a myki card has been shared between multiple people, the dataset will show the movements of those people collectively. It might be argued the data in this case is not 'about' an individual because it does not record the movements of any *particular* individual. In the Deputy Commissioner's view, even where a myki card is shared, each record reveals information about an individual who used the card for a particular journey. The event recorded in the dataset still contains information 'about' that person, even if that individual's identity is unknown or unknowable.

95. The dataset contains information about individuals: namely the location of people at specific times as they started or completed a public transport trip. It also contains more information that can be inferred about those people, for example, their typical public transport movement patterns.

---

[31] *Jurecek v Director, Transport Safety Victoria* [2016] VSC 285 [78].
[32] 'Information and opinions can have multiple subject matters': *Privacy Commissioner v Telstra Corporation Limited* [2017] FCAFC 4 [63].
[33] However, the relevant ticketing rules only permit sharing in certain circumstances, excluding myki cards issued with free travel passes, or with an active myki pass: see PTV, 'Victorian Fares and Ticketing Manual', 1 July 2019, pp 6-7.

Can the identity of the people the information is about be reasonably ascertained from the information?

96. An individual's identity can 'reasonably be ascertained' from information where it is reasonably possible for someone in possession of the information to identify the individual from the information in question. Whether an individual's identity may reasonably be ascertained requires consideration of any potential method of identification, and whether the likelihood, time, effort and reliability of this method is reasonable. This is determined with reference to the information itself, and the context in which it is held or released. The context includes who has access to the information, what other information they are likely to have access to that could be used to link the information, and motivations they may have to re-identify the data. Where there is a reasonable pathway or process for an individual's identity to be ascertained from the information in question, that information will be 'personal information' for the purposes of the PDP Act.

97. In deciding whether an individual's identity can be reasonably ascertained from the dataset, the Deputy Commissioner considered three main issues including the information itself, the context in which the information was released, and potential re-identification scenarios.

98. The information itself is described in paragraphs [80] to [82] above. The dataset is a detailed record of travel movements that used the myki ticketing system over three years. It was subject to some measures to make it difficult to link with any individuals' identity. For example, card numbers were not included as part of the released information, and the times for events were aggregated to the nearest second. Given the nature of the information, there are numerous plausible scenarios in which people might have a strong motivation to attempt to identify the movements of individuals within the dataset. These include:

    - an advertiser seeking to understand where individuals, or groups of people, will typically be at a particular time;
    - a spouse trying to identify unusual trips taken by a partner believed to be having an affair;
    - attempting to locate spouses with the intent to commit family violence;
    - to locate children involved in custody disputes;
    - to aid the commission of criminal acts (stalking, harassment, breaches of intervention orders); and
    - to identify celebrities, or people of notoriety.

99. It is significant the dataset was released to Data Science Melbourne without any restrictions on its use or further dissemination. The dataset was published online for the duration of the Datathon between 24 July 2018 and 26 September 2018. Datathon participants were told they were free to use the dataset in any way they liked, and would not need to sign a non-disclosure agreement. In fact, one Datathon participant republished the dataset online in full, and made it available via their blog from 28 September 2018 until 14 January 2019, when it was taken down voluntarily following inquiries by OVIC. Anyone could access the dataset while it was published online by the Datathon and that Datathon participant, and there were no legal or contractual measures in place that would discourage or prevent people with access to the data from attempting to link it to people's identities, or from sharing it further.[34] It is

[34] The organisers of the Datathon advised OVIC they verbally told Datathon participants that 'under no circumstances should the data be attempted to be re-identified as it was an offence.' This warning was made with reference to the *Privacy Amendment (Re-identification Offence) Bill 2016*, which proposed to criminalise

reasonable to assume the dataset is still held by some Datathon participants and that those participants could potentially disclose the dataset to others.

100. Finally, considering identification scenarios, Data61's analysis, and the University of Melbourne report, present a number of paths by which certain individuals could be identified. These are detailed above at paragraphs [85] and [90]. Some of these scenarios could be accomplished by a malicious actor without access to extensive expertise and computing resources. The Deputy Commissioner considers these represent reasonable pathways to identify individuals. However, even if only sophisticated re-identification methods were available, the Deputy Commissioner considers it a reasonable assumption that some people with access to those sophisticated techniques would attempt to do so, given the high value of the information and its wide dissemination. This includes people who are data practitioners and data enthusiasts.

101. PTV submitted it is necessary to rely on material outside the dataset to identify individuals, including information that is not publicly available or generally known. In the Deputy Commissioner's view, this does not preclude a finding that the dataset contains personal information. Rather, it is a relevant factor that must be considered as part of an evaluative assessment of whether the information is personal information.

102. PTV also submitted it was not possible to identify individuals with 'certainty', and that the process required to re-identify individuals went beyond what is 'reasonable'. However, based on the expert opinions of the Data61 analysts and the University of Melbourne academics, the Deputy Commissioner considered it is possible to re-identify individuals with a high degree of certainty, and that the process required to do so does not go beyond what is reasonable.

103. Having considered these matters, and all the other submissions made by PTV in this investigation, the Deputy Commissioner is of the opinion that the identity of a substantial proportion of the individuals whose travel movements are recorded in the dataset can reasonably be ascertained.

Conclusion

104. PTV put forward several reasons why, in its view, the dataset did not contain personal information. In the Deputy Commissioner's view, the approach suggested by PTV was a literal and technical approach that has been warned against by authorities in discussing the definition of personal information. A literal and technical approach does not support the objects of the PDP Act, which is intended as beneficial legislation.

105. The evidence before the Deputy Commissioner suggests the identities of individuals can be extracted from the dataset with relative ease. PTV has provided no persuasive evidence to the contrary, and has instead relied on technical arguments about the definition of personal information. The facts before the Deputy Commission show that the dataset contains a wealth of information about the travel movements of Victorians, which was disclosed with no effective controls in place to guard against re-identification.

---

the re-identification of de-identified datasets released by Australian Government entities. However, that Bill has not to date passed and would not generally apply to Victorian Government datasets such as the myki dataset. The verbal warning provided by the Datathon organisers, although well intentioned, was inaccurate, and contradicted written guidance provided on the Datathon website that said the data could be used freely.

106. The Deputy Commissioner found the dataset contains information about people whose identity can reasonably be ascertained. This information is personal information and must be handled in accordance with the IPPs in the PDP Act.

107. It is now necessary to consider whether the use or disclosure of the personal information by PTV, when it shared the dataset with Data Science Melbourne for the purpose of the Datathon, was permitted by IPP 2.

Was the disclosure of personal information by Public Transport Victoria permitted by Information Privacy Principle 2?

108. IPP 2 relates to the use and disclosure of personal information. IPP 2 provides that personal information collected for one purpose (the primary purpose) must not be used for any other purpose (a secondary purpose) unless an exception applies. Exceptions include:

- where the secondary purpose is related to the primary purpose, and the use or disclosure would reasonably be expected by the individual the information is about;[35]
- where the individual has consented to the use or disclosure;[36] or
- in other specific circumstances, for example, for public interest research, where disclosure is required by law, to lessen or prevent a serious risk to health, safety or welfare, or necessary for certain law enforcement purposes.[37]

109. The Deputy Commissioner found PTV disclosed the myki dataset to Data Science Melbourne in or around 12 July 2018. As noted previously in this report, PTV does not consider the myki dataset contains personal information. As such, PTV's view is that there has been no breach of the IPPs as a result of disclosing the dataset to the Data Science Melbourne.

What was the purpose of collection?

110. In submissions to OVIC, PTV described the purpose for which it collected the information as follows:

> *The information collected is for a lawful purpose and is necessary for ticketing functions, including to calculate the correct fare, public revenue and cost recovery in provision public transport services, provide reduced fare for eligible customers, to verify requests for refund where/if a customer is charged incorrectly and/or disputes a charge, verify ongoing entitlements, compliance and enforcement. A customer who undertakes a journey in a passenger vehicle, or makes an entry to a compulsory ticket area, for which a fare is required, must pay at least the correct fare in accordance with the conditions contained in this manual for the travel in a passenger vehicle that consists of or includes the journey or for the entry.*
>
> *Also, this data is collected to understand, diagnose and to support data driven decision making around the public transport network. This is the only source of this information. …*

111. The PTV myki privacy policy states:

> *The primary purpose for which PTV collects myki ticketing data is to facilitate the provision and operation of the myki ticketing system, in accordance with PTV's functions under the Transport Integration Act 2010.*

[35] IPP 2.1(a).
[36] IPP 2.1(b).
[37] IPP 2.1(c) – (h).

…

Information is collected to understand, diagnose and to support data driven decision making around the public transport network including:

- *calculate the correct fare, public revenue and cost recovery in provision public transport services*
- *provide reduced fare for eligible customers*
- *verify requests for refund where/if a customer is charged incorrectly and/or disputes a charge, verify ongoing entitlements*
- *ticketing compliance and enforcement*
- *planning, including safety and security, for public transport strategies and investments*
- *patronage trends and understanding how people move around the network*
- *impact to customers at station/stops during major occupation works or disruptions and communications*
- *crowd flow management during major events for safety purposes*
- *identifying cards which require compensation due to an unforeseen event on the network*
- *insights to communication and education campaign analysis such as auto top up campaign tracking*
- *insights to understanding of customers to improve/tailor campaigns accordingly*
- *monitoring new products/devices e.g. Mobile myki or Quick Top Up enquiry machines.*[38]

112. Having considered PTV's submissions and the myki privacy policy, the Deputy Commissioner is of the opinion that the primary purpose for which PTV collects personal information is to facilitate the provision and operation of the myki ticketing system. Part of operating the system is understanding, diagnosing and supporting data driven decisions about the public transport network as outlined in the privacy policy.

What was the purpose of disclosure?

113. PTV informed OVIC the purpose for which the dataset was disclosed to Data Science Melbourne was in response to a request from DPC, and to support a Datathon event:

in which teams of data science students and professionals work with data sets to both:

- *Provide actionable insights to assist government decision making relating to a real-world issues; and*
- *Provide new insights in the data.*[39]

114. The Melbourne Datathon website describes the purpose of the Datathon:[40]

- *To learn from each other and cross pollinate skill sets*

[38] Public Transport Victoria website. 'Collection of personal information', *myki Privacy Policy*. https://www.ptv.vic.gov.au/footer/legal-and-policies/myki-privacy-policy/.
[39] This description of the event and its purpose was included in an email from PTV and in the PTV privacy impact assessment for the data release.
[40] Data Science Melbourne website. 'Why?', *Melbourne Datathon 2018 website.* http://www.datasciencemelbourne.com/datathon/.

- *To provide a stage for potential employers and employees to meet*
- *To create a buzz in Melbourne around Data Science and reverse the brain drain*
- *To solve a real world problem that could impact the lives of all Australians*
- *To have fun!*

115. The Datathon's website also says that, while there are no set tasks for the participants:

> *As a true 'data explorer', you will have to come up with your own questions for the data. We want the datathon to be just like a real data science consulting task. Ask yourself what the data provider might want to learn, and how you might go about presenting that.*

116. Based on PTV's submissions, the content of the PIA, and public descriptions of the Datathon event, the Deputy Commissioner considers the purpose of the disclosure of the dataset to the Datathon was to support the Victorian data science community and economy. PTV may also have hoped to receive insights from the Datathon participants that might have been useful for it in operating the public transport system, but this appears to have been secondary to the altruistic objective of supporting the Victorian data community through the Datathon.

117. This purpose does not match the primary purpose of collection. As such, to be authorised by IPP 2, an exception in IPP 2 must apply.

Does an exception in Information Privacy Principle 2 apply?

118. As the purpose of collection (the primary purpose) is different to the purpose of disclosure (a secondary purpose), it is prohibited by the PDP Act unless one of the exceptions in IPP 2 applies. This section considers the most relevant exceptions to IPP 2 and whether they apply.

Information Privacy Principle 2.1(a) – related secondary purpose that is within reasonable expectations

119. The most relevant exception to IPP 2 is IPP 2.1(a), which permits disclosure of personal information:

- for a secondary purpose that is related to the primary purpose; and
- where that use or disclosure would reasonably be expected by the individuals the information is about.

120. The Deputy Commissioner is satisfied the secondary purpose is related to the primary purpose. Although the purpose of the disclosure was mainly altruistic, PTV did hope to obtain insights into the data that might have assisted its operation of Victoria's public transport network. This purpose is related to the primary purpose of collection as described at paragraph [111] above, which includes data analytics to improve public transport. The Deputy Commissioner considers the relation between these two purposes is somewhat remote, and could not be described as 'directly related', as would be required if the dataset contained 'sensitive information' as defined in Schedule 1 of the PDP Act.

121. A purpose of collection related to the *primary* purpose of collection is not necessarily permitted by IPP 2. The disclosure must also be reasonably expected by the people whose information was disclosed – that is, Victorian public transport users.

122. In certain circumstances and subject to appropriate controls, an individual might reasonably expect PTV would use and disclose historical myki data for secondary purposes, for example, to better understand the dynamics of the public transport network, or to support planning and investment decisions. The myki privacy policy states 'PTV and its contractors use/disclose personal information for managing and improving public transport ticketing and supporting

products and services.'[41] However, it also says '[i]rrespective of whether your Personal Information or Health Information is stored electronically or in hard copy form, PTV will take reasonable steps to protect it from misuse and loss and unauthorised access, modification or disclosure.'[42] Individuals could reasonably expect that, when disclosed by PTV, personal information would be protected. Further, they could reasonably expect any use or disclosure would be limited to what is required to achieve the primary purpose of improving the public transport network.

123. Releasing a myki dataset to a Datathon would not reasonably be expected by the people the information was about – the Victorian public. This is especially the case for a release that did not involve any limitations or restrictions being applied to potential uses or downstream disclosures of the dataset. This means that IPP 2.1(a) did not permit the disclosure of the dataset to the Datathon.

Information Privacy Principle 2.1(c) – necessary for research or statistics

124. IPP 2.1(c) provides a specific exception to not using information for the primary purpose of collection in situations where personal information is necessary for research or for the compilation or analysis of statistical information. This exception applies where three requirements are met:

- The research is in the public interest;
- The information is not for publication in a form that identifies any particular individual; and
- It is impracticable for the organisation to seek the individual's consent before the use or disclosure.

125. In the case of a disclosure, the organisation must also reasonably believe the recipient of the information will not disclose the information.

126. PTV cannot rely on this exception for two reasons.

127. First, it is unlikely the Datathon could be described as 'research'. The term 'research' is not defined in the PDP Act. The Deputy Commissioner considered the word should be given its usual meaning: 'diligent and systematic inquiry or investigation into a subject in order to discover facts or principles'.[43] The Datathon's purpose appears primarily to develop skills and relationships in the Melbourne data science community, not to conduct research. Although individual Datathon participants sought to discover facts or principles from the dataset, the Datathon was neither 'systematic' nor did its primary purpose appear to be the discovery of facts or principles.

128. The second reason is that PTV had no basis to reasonably believe the recipients of the information would not disclose the information. In fact, the Datathon wrote to PTV and DPC to confirm Datathon participants would be free to use the data without limitation or restriction. PTV was, or should have been, aware there were no restrictions on how Datathon participants could use the data, including on-disclosing it.

[41] 'Use and disclosure of personal information', *myki Privacy Policy.*
[42] 'Data Security and Destruction', *myki Privacy Policy*.
[43] Definition of 'research', Macquarie Dictionary (2017).

Conclusion

129. The Deputy Commissioner found neither IPP 2.1(a), 2.1(c), nor any other exception to IPP 2 permitted the disclosure of the personal information contained in the dataset. In disclosing the dataset to Data Science Melbourne on or around 12 July 2018, PTV contravened IPP 2.1 and therefore interfered with the privacy of the individuals whose personal information was contained in the dataset.

Events leading to the data release and Information Privacy Principle 4.1

130. This section considers whether PTV met its obligation under IPP 4.1 in the lead up to the release of the dataset to Data Science Melbourne to take reasonable steps to protect personal information it holds about myki users. It outlines the events and factors the Deputy Commissioner's investigation identified as leading to the release of the dataset.

What does Information Privacy Principle 4.1 require?

'Reasonable steps' to protect personal information

131. IPP 4.1 requires organisations to take reasonable steps to protect the personal information they hold from misuse and loss and from unauthorised access, modification or disclosure.

132. What reasonable steps are required depends on a wide range of factors, including the nature of the information and how it is held. Organisations must select security measures and controls appropriate to their circumstances and the risks they have to manage. These measures and controls must be proportionate to the potential harm that may result from a failure to protect the information. Factors relevant to assessing what steps are reasonable include:

- the potential impact of a privacy breach (on the people the information is about);
- the type and amount of information; and
- the nature of the organisation, including its size and the resources at its disposal.[44]

What information was Public Transport Victoria required to protect?

133. PTV provided the myki dataset to Data Science Melbourne on the basis the information released did not contain personal information. It is the data this information was derived from, which PTV held and continues to hold, that it is required to protect.

134. The Datathon dataset was extracted from a data warehouse called myki Mirror. The data warehouse contained the trip information that was used to build the dataset, and it also contained information about people who had registered myki cards. This information can be used by PTV to connect particular cards (and associated journeys) with named individuals (with registered myki cards).

135. If PTV were able to modify the dataset so it no longer contained personal information (that is, if the dataset was successfully de-identified), it would not need to handle the modified dataset in accordance with the IPPs. The information this dataset was derived from was based on personal information. Where an organisation proposes to release de-identified data, it is still obliged to take reasonable steps to protect the personal information the dataset is derived from. This includes taking appropriate de-identification measures and applying appropriate risk management and decision-making processes to ensure source data is not subject to unintentional disclosure.

[44] Commissioner for Privacy and Data Protection, 'Guidelines to protecting the security of personal information: 'Reasonable steps' under Information Privacy Principle 4.1' (January 2017), pp 14–15.

What events or factors contributed to the data release?

136. The disclosure of the myki dataset to Data Science Melbourne created a risk that people could discover information about the travel movements of Victorians. This information was derived from personal information held by PTV that it is required to take reasonable steps to protect.

137. To consider whether PTV took reasonable steps to protect the personal information it held, the Deputy Commissioner considered the events and factors that led to the data release decisions, as identified by the investigation. These events and factors are examined to consider whether any of them point to a failure on PTV's part to take reasonable steps to protect the personal information it held.

Reliance on flawed privacy impact assessment

138. A PIA is a systematic assessment process that seeks to help organisations identify the impact a program or activity might have on individuals' privacy. The PIA sets out recommendations for managing, minimising or eliminating that impact.

139. PTV completed a PIA before the data release and relied on it when deciding to release the information to Data Science Melbourne. PTV said:

> *A PIA was undertaken to assess how myki data could be released in a way that supported meaningful analysis [and] was undertaken at the request of DPC … PTV's privacy team and the data owner were consulted over the release of the data and what was required beyond completing and signing off a PIA. No other actions were required given that the data was not identified as personal information within the meaning of the Privacy and Data Protection Act 2014.*

140. The PIA was the only formal or documented record of decision by PTV to release the dataset to Data Science Melbourne.

141. The PIA, dated 17 January 2018, was recorded in a PIA template. The template used by PTV in developing its PIA was issued by the Office of the Commissioner for Privacy and Data Protection (**CPDP**), OVIC's predecessor organisation.[45]

142. To understand PTV's PIA, it is necessary to describe the template. The template document consists of four parts. Part 1 of the template asks users to describe the program or project to which the PIA relates, and identify the types of information that will be handled. This is to determine the scope of the privacy analysis required. The template instructs users that, should Part 1 conclude no personal information will be handled, the PIA process is complete and Part 4, the PIA summary and sign off, can be completed. Parts 2 and 3 assess compliance with the IPPs and require the template's user to identify privacy risks and possible risk mitigation strategies. The template indicates Part 2 and Part 3 do not need to be completed if Part 1 concludes no personal information is involved.

143. PTV's PIA describes the proposal being assessed as:

> *A 'hackathon' event sponsored by the Department of Premier and Cabinet and Transport for Victoria in which teams of data science students and professionals work with data sets to both*
>
> - *provide actionable insights to assist government decision making relating to a real-world issue; and*
> - *provide new insights into the data*

[45] Commissioner for Privacy and Data Protection, *Privacy Impact Assessment Template* (May 2015).

The myki data is owned by Public Transport Victoria. The data will be extracted from PTV's Data Analytic Platform. For the duration of the hackathon the data will be hosted on an analytic platform provided by a third party contracted by Department of Premier and Cabinet.

144. The PIA states no personal information will be involved in the myki data release. Under the heading '3.1 Personal Information', the template asks users to 'Please list or attach as an appendix, all the personal information the program will collect, use or disclose.' Below this heading PTV has written:

No personal information capable of identifying [an] individual will be disclosed.

145. No analysis or reasoning for this conclusion is provided.

146. However, the template does ask users to consider the risk of re-identifiable information. Under the heading '3.5 Re-identifiable Information', the template provides the following guidance:

Many programs rely on the use of de-identified or non-identifiable information. When such information is used it needs to be treated with caution and afforded many of the same privacy protections as personal information, where there is a potential for re-identification to occur. This is particularly the case where a program involves data matching/linking activities. For that reason, when assessing privacy of personal information, potentially re-identifiable information should be protected in the same way as personal information.

147. The template then asks whether the program will collect, use or disclose re-identifiable information. PTV responded as follows:

No. There is no way to link the public transport travel patterns of individual mykis to specific people via the encrypted internal card ID – this is not publicly available and will be encrypted in any case. The only remaining risk is that someone may attempt to identify a specific myki card based on the travel patterns but this would require a detailed knowledge of when and where a person had used public transport – basically a travel diary – and it would be very difficult to distinguish from other cards with similar travel patterns. In the unlikely event that this succeeded it would only reveal which Public Transport modes and stops the card had appeared at.

148. The analysis in Part 2 and Part 3 of PTV's PIA is incomplete, as the PIA concluded the program contained no personal information or re-identifiable information. This means the portions of the PIA template designed to assess adherence to the IPPs and to be populated with potential privacy risks and remediation is unpopulated.[46]

149. The PIA was completed by a PTV data scientist and signed off by the data owner and PTV Chief Information Officer. It was then sent to PTV's governance and legal team by the data scientist. In a response email, PTV governance and legal told the data scientist:

Given that there is no personal or confidential information involved in this project, privacy or confidentiality laws are not applicable to this project.

If any changes to the scope, objectives or information particularly if personal of confidential information is involved in the future you must complete the PIA / risk assessment and contact the Privacy Team to discuss.

150. Email correspondence between the PTV data scientist and DPC show both the PIA and the email from Governance and Legal were taken to be 'advice' which formed the basis of an 'OK to release the myki data to the hackathon.'

[46] Privacy Impact Assessment conducted by PTV, 'Table 6: Risk Mitigation'.

151. The Deputy Commissioner finds the PIA process undertaken by PTV was flawed in a number of respects.

152. First, there are factual inaccuracies in how the Datathon is described in the PIA (see extract at paragraph [143] above). The PIA states the information will be held on an 'analytics platform hosted by a third party contracted by DPC.' While there was a contractual relationship between DPC and Data Science Melbourne, it was a funding arrangement that included no requirements about the protection of data. Further, the data was not hosted on an analytics platform provided by the Datathon organisers. Extracts of the data were given directly to participants as a csv file. Later, Data Science Melbourne provided Datathon participants with links to the full dataset. These links were made publicly available on the internet. These were important contextual factors for the data release that should have been included in any PIA.

153. Second, the PIA incorrectly concluded no personal information would be disclosed. The dataset as released to Data Science Melbourne contained personal information (see paragraphs [74] – [106]). The PIA briefly referred to, but too quickly dismissed, the risk of re-identification in the text quoted at paragraph [147] above. The PIA does not record any considered analysis of the risk of re-identification.

154. Third, the scope of the PIA was too narrow. It considered only the de-identified dataset that would be disclosed to Data Science Melbourne and not the original source dataset. A PIA should consider *uses* of personal information by an organisation. In this case, PTV was using information it held, which was personal information, to create a de-identified dataset for use in the Datathon. This use of information should have raised privacy questions the completed PIA should have considered. For what purposes does PTV collect myki travel information? Does that purpose permit its use for activities such as creating a de-identified dataset for the Datathon? How would public transport users expect PTV to use information about them collected through the myki system? If PTV had correctly considered that, even if de-identification was successful, it would be using personal information to create that de-identified data, the rest of the analysis in the PIA template may have been completed. This may have highlighted some of the risks listed in this report.

155. Fourth, the PIA was conducted at a single point in time. It was not reconsidered as the details of the project changed. PTV's governance and legal team advised the PTV data owner a further PIA should occur if there were changes to the scope, objectives, or information associated with the project.[47] However, this did not occur even when substantial changes to the data being released were negotiated and made, for example, when the amount of data being provided expanded from one to three years, or when information linking trips to categories of myki concession cards was included with the dataset.

156. Finally, the PIA template was used to achieve more than it was designed to do: the PIA was used by PTV as the authorising document or 'sign-off' for the data release. However, the PIA template is designed to help organisations identify and treat (manage) privacy risks. It is not designed as a complete record of decision for data release approvals, and as such does not touch on non-privacy considerations relevant to a data release decision, for example, questions about the utility of the data for its intended use, or ownership and licencing issues.

157. Despite these issues, the Deputy Commissioner found that PTV and DPC placed significant weight on the PIA. The email chain referred to above at paragraphs [149] and [150] show the

[47] See paragraph [149], above.

PIA created a false sense of security. This false sense continued into the post-incident response conducted by PTV and DPC, where the PIA was relied on as confirmation the dataset had been de-identified. It appears PTV and DPC decision makers assumed that because a PIA had been written, privacy risks had been addressed. In fact, the flawed threshold assessment in the PIA and its overly narrow scope meant privacy risks were not adequately considered. Further, the failure to conduct a new PIA when the details of the project changed meant there were lost opportunities for privacy risks to be remedied at a later time.

158. The flawed approach taken by PTV in completing the PIA, and the resultant confidence it provided to PTV and DPC that privacy issues had been considered and addressed, are key contributors to this incident.

Inadequate de-identification measures

159. PTV took some steps to modify the data to prevent individuals from being re-identified. The main step was to replace PTV's internal card ID number with a different generated value. This transformation was achieved by applying a relatively simple algorithm to the original PTV identifier.

160. The way the internal card key was generated, and the way it was used to create the released identifier, resulted in the identifier containing several characteristics that may be able to be used to support an attack against the data. This opens the dataset up to additional re-identification attacks based on the identifier number.

161. The Deputy Commissioner found that the algorithm used, although simple, is not demonstrably reversible. PTV has also noted 'even if de-randomised, the internal card key cannot be related to any publicly available data – to do that the myki back end system would have to be compromised.' However, the method used does result in patterns in the data which may support re-identification attacks.

162. When generating identifying numbers to link de-identified datasets, the Deputy Commissioner considers it is better practice to either:

- use a random number, with no connection to an agency's internal identifier. This is the most secure method, and would be appropriate if it is not necessary to link the dataset back to the original source data or to a later data release (as appears to have been the case in this data release); or
- generate a meaningless number using an industry standard, secure hashing algorithm. A hashing algorithm takes a string of text and converts it into another, seemingly random, string of text, in a way that is effectively irreversible. A hashing algorithm can be regarded as secure when its details are published and have been subject to academic and industry scrutiny, with no weaknesses identified. This method is slightly less secure than applying a truly random number, but it adds utility by allowing future releases of additional records that use the same generated number.

163. PTV has not identified any other steps taken for the purpose of making the dataset less identifiable. A range of additional techniques might have been, but were not, considered to treat the data. Examples include the following:[48]

- **Sampling** – providing access to only a fraction of the total existing records, thereby

[48] Examples drawn from Office of the Australian Information Commissioner, 'Data modification and data reduction techniques', *De-identification and the Privacy Act*, March 2018.

creating uncertainty that any particular person is included in the dataset.

- **Choice of variables** – removing quasi-identifiers that are unique, or which in combination with other information are reasonably likely to identify an individual. Examples of quasi-identifiers that might have been removed by PTV are the 'card type' records that had very few linked cards.
- **Rounding** – combining information or data likely to enable identification of an individual into categories. This may have included by rounding the time of travel events to less granular times than the one second intervals that were published.
- **Perturbation** – altering information that is likely to enable identification in a small way, such that aggregate data is not significantly affected, but the original values cannot be known with certainty.
- **Swapping** – swapping information that could enable the identification of an individual for one person with the information for another person with similar characteristics to hide the uniqueness of some information.
- **Manufacturing synthetic data** – creating new values generated from original data so overall totals, values and patterns are preserved, but do not relate to any particular individual.

164. PTV did not seek external expertise to assist with de-identifying the dataset prior or during the dataset's release for use in the Datathon. At the time PTV was considering releasing the dataset, it had no documented policies or procedures for de-identification of data. The flaws in the release process highlight the need for relevant expertise or appropriate policies to support de-identification processes. The Data Vic Access Policy Guidelines state agencies should have a policy to ensure the correct de-identification of data if information based on personal information will be released publicly:

> *To ensure that datasets containing personal, health and/or confidential information are correctly and consistently de-identified and or aggregated in order to be made available under the Policy, a formal procedure must be documented and adhered to by agencies.*[49]

165. The Deputy Commissioner finds that, in preparing the dataset for release, PTV did not take a methodical approach to de-identifying the released data, let alone achieve best-practice. Assuming PTV intended to release the information as 'open data', it is also noteworthy that the approach it took did not adhere to the Data Vic Access Policy Guidelines.

Over reliance on safety of the data, at the expense of other 'safes'

166. One model for managing the risks associated with data sharing and access decisions is the 'Five Safes' framework. In Australia, the five safes framework has been promoted by the Australian Bureau of Statistics and others.[50] Each 'safe' refers to an independent but related aspect of disclosure risk. The framework is designed to facilitate safe data release. To do this, it poses specific questions to help assess and describe each risk aspect (or safe) in a qualitative way. This allows data custodians to place appropriate controls, not just on the data itself, but also on the manner in which data are accessed.

---

[49] Data Vic Access Policy Guidelines for the Victorian Public Sector (Version 2.1), November 2016 [5.2].
[50] See, e.g., Australian Institute of Health and Welfare, 'The Five Safes Framework', <https://www.aihw.gov.au/about-our-data/data-governance/the-five-safes-framework>. Australian Bureau of Statistics, 'Managing the Risks of Disclosure: the Five Safes Framework', <http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/1160.0Main%20Features4Aug%202017>.

167. Although neither PTV nor DPC are required to use this framework, and it has no formal status in Victoria, it is a useful model to consider the different ways risks of disclosure or re-identification could have been managed.

168. The five elements of the framework are:

- Safe People: Can the users be trusted to use it in an appropriate manner?
- Safe Projects: Is this use of the data appropriate?
- Safe Settings: Does the access facility limit unauthorised use?
- Safe Data: Is there a disclosure risk in the data itself?
- Safe Outputs: Are the statistical results non-disclosive?

169. In this context, the only 'safe' that appears to have been considered was 'safe data'. The only question PTV appears to have considered in deciding whether to release the dataset was whether it contained personal information.[51]

170. Employees of DPC and PTV involved in organising the Datathon were aware there were no protections on the data beyond the de-identification measures applied to it. An email dated 16 May 2018 sent by the Datathon organisers to PTV, and forwarded to DPC, asked:

Can you also confirm that there are no restrictions on the data being released to us? The type of things I mean is would we be able to put it in a cloud service so the participants can access it easier, rather than handing it out on USB sticks? Would the participants be able to build web apps if they wanted to, which would mean the data would have to be stored somewhere in the cloud? Is the data to be only used for the Datathon, or are they free to use it for whatever they want to after the event?

171. There was a contract between Data Science Melbourne and DPC, but it only discussed the sponsorship DPC was providing to the Datathon. It did not include provisions about information security or the protection of the data.[52]

172. PTV relied exclusively on de-identification of the data to manage the risk of people attempting to re-identify people in the dataset. By overlooking other possible means of protecting the information, PTV increased the risk of the data being re-identified. Additional considerations PTV could have completed in the Five Safes framework include the following:

- Limiting the disclosure of the dataset by the Datathon to a known and fixed list of Datathon participants;
- Ensuring those participants were subject to contractual or legal obligations not to attempt to re-identify the data, not to on-disclose the data, and to destroy the data at the conclusion of the Datathon;
- Ensuring the data was held on a known and secure system that limited the possibility of data being extracted and retained by Datathon participants after the Datathon concluded.

[51] The 'Five Safes' framework works where all safes are able to be considered. Where data is made available on the Internet, as 'open data', the access facility *Safe Settings* control is not available, and the *Safe People* control cannot be applied. Therefore it is not appropriate to 'open data' unless restrictions on downstream use or export can be controlled.

[52] Victorian Common Funding Agreement between DPC and Data Science Melbourne, D18/107723, 26 June 2018.

173. PTV's sole reliance on its assumption the dataset was anonymised or de-identified, especially considering the issues with the de-identification approach it took, was one of the factors leading to the exposure of personal information.

Lack of clarity about division of responsibilities between Public Transport Victoria and the Department of Premier and Cabinet

174. PTV advised OVIC that, at the time it was working with Data Science Melbourne to prepare the data, it 'understood that an appropriate governance process was already in place, including Data Science Melbourne [having signed a] confidentiality deed with DPC'. PTV said 'DPC managed the Datathon and had the relationship with Data Science Melbourne regarding the use, handling, storage or release of data provided for the Datathon. PTV 'understood Data Science Melbourne to be acting on behalf of DPC.' During the investigation, PTV staff described Data Science Melbourne as a 'DPC contractor' and indicated frustration at not having known how Data Science Melbourne was intending to use and disclose the information in question.

175. On the other hand, DPC told OVIC it had held no discussions with Data Science Melbourne and PTV about the use, handling, storage or release of the data. The only contractual relationship between DPC and Data Science Melbourne was a sponsorship arrangement for the Datathon which imposed no requirements relating to the use of the data provided by the Victorian public sector. DPC said that 'PTV made the decision to release the data. DPC [does] not have a role in approving data releases by other departments or agencies.' DPC understood PTV had worked out the details of the data release, and that all privacy and other risk assessments were undertaken by PTV.

176. PTV misunderstood the relationship between DPC and Data Science Melbourne. This is clear from the above PTV and DPC comments, as well as how DPC's relationship with Data Science Melbourne was described in the PIA. There was a lack of clarity between DPC and PTV about who was responsible for protecting the data and considering any privacy risks.

177. When the dataset was released for use in the Datathon, PTV appears to have understood DPC had a role in overseeing Data Science Melbourne's use of the data, and that Data Science Melbourne was a contractor receiving and handling the data on DPC's behalf.

178. The lack of a shared understanding (at the time of the data release) about who was responsible for managing these risks likely contributed to the other issues identified in this section, and in turn, contributed to the incident.

Did Public Transport Victoria fail to take reasonable steps to protect the personal information?

179. The information from which the myki dataset was derived contained detailed travel records of millions of Victorians. This information does not fall within the definition of 'sensitive information' referred to in the PDP Act. However, it is information that can be regarded as delicate;[53] the people who the information is about would likely expect it would be subject to a high degree of protection. The scenarios in which someone may be motivated to re-identify the dataset indicate the potential and foreseeable risks of harm from unauthorised disclosure

[53] See Office of the Victorian Privacy Commissioner, *Guidelines to the Information Privacy Principles* (2011) 14-15.

or access to travel movement information.[54] As such, PTV should have taken significant steps to protect this information.

180. To assure itself the personal information it held was protected during the Datathon, PTV conducted a flawed PIA process, and made inadequate modifications to the data released. In light of the nature of the information in question, and the amount involved, this was insufficient.

181. During the investigation, PTV repeatedly claimed that by completing a PIA, it followed the process that was put in place at the time by OVIC's predecessor, CPDP. However, the flaws in PTV's completed PIA were such that it is not accurate to say the PIA was completed in accordance with guidance issued by CPDP. PTV also overlooked a range of other guidance material released by OVIC, and its predecessor regulators, about de-identification that would have assisted it to identify the privacy risks to the dataset, and better de-identify the data.[55] PTV also failed to adhere to the DataVic Access Policy Guidelines.[56] PTV did not follow an appropriate process when considering the dataset for release.

182. The Deputy Commissioner's view is that the factors and events outlined above indicate a number of failures by PTV to take reasonable steps to protect this information during the data release process. The reasonable steps PTV failed to take include:

- the flaws in the PIA process discussed at paragraphs [138] to [158];
- PTV's over-reliance on the PIA as the sole authorising document for the data release, in the absence of other documented policies or procedures that inform data release decisions;
- an absence of contractual or other controls being imposed on the data recipient (Data Science Melbourne), or down-stream users (including Datathon participants); and
- the inadequate de-identification measures applied to the dataset, and the failure to take a methodical approach to de-identifying the data.

183. PTV contravened IPP 4.1 by failing to take reasonable steps to protect the information it held about the public transport trips of Victorians while considering Data Science Melbourne's request for the information, and while preparing the dataset for release.

[54] See the examples listed at paragraph [98], above.
[55] The OVIC paper, 'Protecting unit-record level personal information', May 2018 specifically cautioned against releasing unit-level record data. See also CPDP, 'De-identification Background Paper', 2016.
[56] See para [164] above.

Recommendations

184. PTV breached the PDP Act with respect to information it holds about people who have used the myki system. While this breach is in part due to decisions made by PTV, it is also influenced by wider factors that are relevant to the whole Victorian public sector. As such, this report makes recommendations to PTV and to the Victorian public sector more generally. OVIC also considers better regulatory guidance could have been provided to PTV to support it in completing its PIA, so proposes to enhance its PIA guidance.

Recommendation 1: The Department of Transport to document policies and procedures for data release decisions

185. One of the causes of this incident was apparent uncertainty within PTV about how data release decisions should be made. This was manifested through PTV's over-reliance on its PIA as a decision-making document and the lack of clarity between PTV and DPC about their mutual responsibilities. This resulted in PTV not considering protections that could be applied to its data once given to Data Science Melbourne.

186. OVIC recommends PTV document policies and procedures that make clear to internal PTV stakeholders how data release decision should be made, what considerations should be taken into account, who should make the decisions and how decisions should be documented.

187. **Specified Action 1**: The Department of Transport must develop and document policies and procedures for data release decisions and provide OVIC with a copy of these policies and procedures by 1 March 2020. The policies and procedures must:

- clearly explain how data release decisions should be made including identifying:
  - the considerations that must be taken into account;
  - who is authorised to make decisions; and
  - how those decisions should be documented (including through privacy impact assessments);
- consider information data security controls relevant to downstream disclosures of data, scope creep, and data re-identification risks; and
- comply with the PDP Act and Information Privacy Principles.

Recommendation 2: The Department of Transport to continue the rollout of its data governance program initiated by Public Transport Victoria

188. PTV commenced the establishment of a formal data governance program in or around April 2018, following the appointment of an Enterprise Information Management General Manager. Since then, PTV has worked to improve its data governance, for example, PTV has developed a range of documentation including policies and guidance to assist data owners. As part of this program, PTV has delivered training to data owners that touches on privacy and de-identification risks.

189. It appears that if PTV's data governance framework had been in place at the time of the initial discussions that led to the data release, a different result may have occurred. OVIC is satisfied PTV's data governance program goes a substantial way towards addressing the concerns identified in this investigation.

190. **Specified Action 2:** The Department of Transport must implement a data governance program by October 2020. The data governance program must include data governance policies and procedures that are consistent with the PDP Act and Information Privacy Principles.

Recommendation 3: Training

191. The PIA conducted by PTV was flawed. Training should be provided to all Department of Transport data owners about how to identify privacy risks in new projects and on how to complete a PIA.

192. Training should also be provided to all executives and data owners that may be involved in collection, management, or release of data. This should be conducted consistent with the governance framework identified in Specified Action 2 above.

193. **Specified Action 3:** The Department of Transport must deliver training about the above data release policies and procedures and data governance policies and procedures to all relevant staff and data owners in the Department. The training must enable relevant staff and data owners to identify risks in the Department's operations, match those risks to policies and procedures (including those developed as part of specified action 1 and specified action 2), and give effect to those policies and procedures.

- Delivery of training on data release policies and procedures must commence no later than 1 April 2020.
- Delivery of training on data governance policies and procedures must commence no later than 1 November 2020.
- The Department must provide OVIC with training course materials and a schedule for training when training commences.

Recommendation 4: Reporting

194. To provide assurance to OVIC that these Specified Actions are being implemented, the Department of Transport will provide a report on progress against the three Specified Actions and the Department's ongoing commitment to them.

195. **Specified Action 4**: The Organisation will provide OVIC with a report of its progress and compliance with Specified Action 1, Specified Action 2 and Specified Action 3, on 2 March 2020, 1 September 2020 and a final report on 1 March 2021.

Recommendation 5: uplift in data capability across the Victorian Public Sector

196. This incident demonstrates the challenges in identifying privacy risks in large and complex datasets. It is important the Victorian public sector, which possesses many large and sensitive data holdings, have a high level of data literacy.

197. As such, OVIC recommends a training program be developed and delivered over the next two years to increase data literacy at executive levels in the Victorian public service generally. This training program could likely be best developed in consultation with the Victorian Centre for Data Insights, OVIC and other stakeholders.

198. **Recommendation 5**: OVIC recommends the Victorian government deliver training programs to uplift the data capabilities of senior leaders in the Victorian public sector.

Recommendation 6: process to support data release decisions

199. Agencies such as PTV must have internal data governance processes and data expertise. However, it is unlikely all agencies will be able to develop the technical skills necessary to make safe data release decisions. OVIC therefore suggests the Victorian government develop a process to support these data release decisions that involves some degree of oversight from an appropriately resourced and experienced agency.

200. The Victorian government should develop a whole of public sector process for publishing open data where public data includes unit level information relating to individuals or their behaviour. This process could, as a start point, be modelled on the Australian Government's *Process for Publishing Sensitive Unit Record Level Public Data as Open Data*. This process could be developed as part of the ongoing review of the DataVic Access Policy Review being conducted by the Department of Premier and Cabinet.

201. **Recommendation 6**: OVIC recommends that the Victorian government develop a centralised process for publishing public data where that publication includes unit level information.

Recommendation 7: improved Privacy Impact Assessment guidance

202. The PIA completed by PTV was completed using a template document developed by OVIC's predecessor organisation, CPDP. The Deputy Commissioner considers PTV could have been better supported in completing its PIA through greater regulatory guidance. As such, OVIC is committing to enhance its PIA guidance by publishing a guide to sit alongside the PIA template which will provide greater assistance to agencies completing PIAs. As at time of publication of this report OVIC has already done this, based in part on lessons from this investigation.

203. **Recommendation 7:** OVIC recommends it promote its improved PIA guidance to assist agencies in conducting PIAs.

Compliance notice and publication of report

204. As discussed above, under section 8C(2)(e) of the PDP Act, the Deputy Commissioner can issue a compliance notice. A compliance notice may be issued by the Deputy Commissioner under section 78 of the PDP Act in response to a serious, flagrant or repeated breach of the IPPs. A compliance notice requires an organisation to take specified action within a specified period for the purpose of ensuring compliance with the IPPs.

205. An investigation may also lead to the publication of a report and recommendations under section 111 of the PDP Act. Section 111 permits the Information Commissioner to publish a report where they consider it is in the public interest to do so. The Commissioner may report on any act or practice the Commissioner considers to be an interference with privacy, or report about any matter generally relating to the Commissioner's function under the PDP Act.

Decision to issue a compliance notice

206. As noted above, the Deputy Commissioner found that PTV breached the IPPs in the course of releasing the dataset to Data Science Melbourne. On 3 June 2019, the Deputy Commissioner advised PTV of a preliminary view that a compliance notice should be issued in response to the breach, and her reasons for reaching that view.

207. The Deputy Commissioner considered PTV's submissions, as well as all of the other material described in this report, before deciding that PTV's breach was a 'serious' contravention of the IPPs for the purpose of section 78(1)(b)(i) of the PDP Act, and that a compliance notice should be issued. In reaching that view, the Deputy Commissioner considered factors including:

- the type of information in the dataset;
- the amount of information involved, and the number of people to whom it relates;
- the extent of harm to individuals and the likelihood of further harm that may result from the incident;
- the potential impact of the breach on public trust;
- PTV's response to the incident and its conduct during the investigation;
- PTV's willingness to implement the Deputy Commissioner's proposed recommendations;
- PTV's views on the definition of 'personal information' and related matters; and
- the fact that, to the best knowledge of the Deputy Commissioner, this was the only such incident involving PTV and PTV has not previously been subject to regulatory action from OVIC or its predecessors.

208. Although there were factors both for and against issuing a compliance notice, on balance, the Deputy Commissioner decided that a compliance notice under section 78 of the PDP Act should be issued.

209. A copy of the compliance notice is included at **Attachment A**.

Decision to publish a report

210. On 3 June 2019, the Deputy Commissioner wrote to PTV and DPC to advise of OVIC's preliminary view that a report of this investigation should be published by OVIC. She provided reasons for that view.

211. The Information Commissioner considered the Deputy Commissioner's reasons, PTV and DPC's submissions, and the terms of this report of investigation and other supporting material. In deciding whether publishing a report of investigation was in the public interest, the Information Commissioner considered a number of factors including:

- the need to provide transparency to the community about this issue, to allow the community to understand both the issue and the response taken by the public sector;
- the educative value of publishing an investigation report for PTV, DPC, OVIC and other data custodians
- the potential for a public report to lead to better decisions on open data; and
- a consideration the dataset vulnerability was likely to come to wide public attention at some point, and that it was preferable that it do so in the context of a regulatory investigation, and a compliance notice requiring remediation action.

212. The last factor requires further discussion. The Information Commissioner considered publishing a report may raise community awareness of vulnerabilities in the data, leading to re-identification attempts against copies of the data. Copies of the dataset may still be held by Datathon participants, or others who downloaded the dataset when it was available on the Internet. It was for this reason that OVIC did not suggest public notification about the incident when it first became aware of it.[57]

213. A very significant consideration for the Information Commissioner was his view that, regardless of whether OVIC published a report, information about the vulnerability would come to light in any event. Many people within and outside the public sector are aware of the vulnerability to the data, and the University of Melbourne academics advised OVIC they were preparing their own report into the vulnerability. The Information Commissioner considered it highly likely that this incident would come to wide public attention at some point, and that it was preferable that it do so in the context of a regulatory investigation, and a compliance notice requiring remediation action.

214. Although there were factors both for and against publishing a report, on balance, the Information Commissioner decided that it was in the public interest to publish a report. As such, the Information Commissioner published this report under section 111(3) of the PDP Act.

[57] See discussion at paragraph [67] above. As noted above, this risk has declined due to the passage of time.

Department of Transport response

The privacy of people travelling on Victoria's public transport network is taken very seriously and is very important to DoT. DoT will implement the actions set out in the Report and has already started work on a number of them.

*DoT has carefully considered your Report and does not accept the finding that the myki travel information (**data**) disclosed by PTV contains or constitutes 'personal information' or a breach of individuals' privacy. The data is not information 'about an individual whose identity is apparent, or can be reasonably ascertained, from the information' on either broad or narrow interpretation of 'personal information' for the purposes of the Privacy and Data Protection Act 2014 (**the Act**).*

DoT does not accept that the process followed by PTV was 'flawed'. PTV carefully anonymised data before it was disclosed, following the privacy impact assessment (PIA) template and guidelines in place at the time, developed by your Office's predecessor and available on the OVIC website. DoT notes that these guidelines are still current and now subject to review.

PTV took relevant steps to ensure that personal information was not contained in the data, and only raw touch-on and touch-off data was released, with myki card numbers randomised. The data relates to travel conducted in public places, and journeys from a public transport stop to another public transport stop. This is not information of a personal or private nature, such as individuals' names, addresses or contact details.

It is not possible to re-identify individuals based on the data PTV provided alone, nor is it possible to establish a complete picture of a person's trip - people should have confidence their privacy is protected. The data is limited in time and a year of date.

A lot more information and further steps are required from other sources, along with private knowledge, data science expertise and capability, for the scenarios mentioned in the Report to arise. Further, the database does not represent a one-to-one relationship between cards, individuals and travel patterns.

Our position remains that a Compliance Notice was not warranted in the circumstances. However, as our priority is the privacy of people travelling on our network, we will work with OVIC, other government agencies and stakeholders to implement the actions specified in the Compliance Notice.

DoT has established a formal data governance program and continues to work on process improvements. We are developing a new, enhanced privacy and research ethics framework, and will be consulting your Office and other stakeholders on this later this year. We will be applying this new privacy framework that will be reviewed by your Office before any data release in the future.

Department of Premier and Cabinet response

*The Department of Premier and Cabinet (**DPC**) does not accept the Office of the Victorian Information Commissioner's (**OVIC**) findings that release of the myki dataset for the purposes of the 2018 Melbourne Datathon necessarily involved the disclosure of any 'personal information' within the meaning of the Privacy and Data Protection Act 2014 (**PDP Act**).*

The data that was released as part of the Datathon did not contain any individuals' names, addresses or myki card numbers. The data did not contain any details of any person's identity. Instead, to use the data to re-identify an individual's myki card travel history involves multiple steps, including cross-matching the data with information from other sources and private knowledge. In DPC's view, the PDP Act was not intended to operate to protect this type of de-identified information.

DPC also understands that, before deciding to release the data in July 2018, a Privacy Impact Assessment was completed relying on the template and guidelines dated May 2015 and 2011 respectively. DPC notes that OVIC acknowledges PTV could have been better supported in completing its Privacy Impact Assessment through greater regulatory guidance at that time. DPC notes that one of OVIC's recommendations is to assist agencies going forward by improving its own Privacy Impact Assessment guidance.

DPC remains committed to protecting the privacy of Victorians, and continuously improving the approach to data capability and publishing processes, while meeting its responsibilities under the PDP Act. DPC remains committed to realising the significant benefits of open data sources to leverage better outcomes for Victorians. Therefore, DPC intends to work closely with OVIC to implement the recommendations contained in your report.

DPC continues as the custodian of the DataVic Access Policy. While we recognise that not all data is suitable for public release, the purpose of the policy includes enabling public access to government data to support research and promote innovation. As part of its commitment to OVIC's recommendations, DPC is currently reviewing the DataVic Access Policy and is considering consulting with the public on the policy.

In relation to OVIC's recommendation concerning publishing data that includes unit level information, DPC is planning to develop specific guidelines and would seek OVIC's input to address concerns about the publication of such data.

Attachment A

COMPLIANCE NOTICE

Under section 78 of the *Privacy and Data Protection Act 2014 (Vic)*

Office of the Victorian Information Commissioner

To: **Department of Transport** (integrating the Public Transport Development Authority)
1 Spring Street,
Melbourne, Victoria 3000
(the **Organisation**)

I, Rachel Dixon, under sections 8B(1)(a) and 8C(2)(e) of the *Privacy and Data Protection Act 2014 (Vic)* (**PDP Act**), serve this compliance notice under Division 9 of Part 3 of the PDP Act.

1. Background

1.1 In or around July 2018, the Organisation created and disclosed a dataset to a third party. The dataset contained information extracted from the myki electronic ticketing system operated by the Organisation.

1.2 While the dataset was claimed to be deidentified, I am satisfied that information about individuals could be reasonably ascertained. Consequently, I consider that the dataset contained personal information.

1.3 Accordingly, in disclosing the dataset to the third party, I am satisfied that the Organisation contravened:

1.3.1 Information Privacy Principle 2.1 by disclosing personal information for a purpose other than that for which it was collected; and

1.3.2 Information Privacy Principle 4.1 by not taking reasonable steps to protect the personal information it held from misuse and loss and from unauthorised access, modification or disclosure.

1.4 I am also satisfied that the contravention was serious having regard to:

1.4.1 the type and amount of information that was released;

1.4.2 the likelihood and extent of harm to individuals that could arise; and

1.4.3 the circumstances surrounding the release.

2. Action and Period Specified

2.1 In accordance with section 78(2) of the PDP Act, this compliance notice requires the Organisation to take this specified action within the specified period for the purpose of ensuring compliance with Information Privacy Principles 2.1 and 4.1.

***Specified Action 1 – Data release policies and procedures*:** The Organisation must develop and document policies and procedures for data release decisions, and provide OVIC with a copy of these policies and procedures by 1 March 2020. The policies and procedures must:

- clearly explain how data release decisions should be made including identifying:
  - the considerations that must be taken into account;
  - who is authorised to make decisions; and
  - how those decisions should be documented (including through privacy impact assessments);

- consider information data security controls relevant to downstream disclosures of data, scope creep, and data re-identification risks; and
- comply with the PDP Act and Information Privacy Principles.

***Specified Action 2 – Data governance program*:** The Organisation must implement a data governance program by October 2020. The data governance program must include data governance policies and procedures that are consistent with the PDP Act and Information Privacy Principles.

***Specified Action 3 - Training*:** The Organisation must deliver training about the above data release and data governance policies and procedures to all relevant staff and data owners in the Organisation. The training must enable relevant staff and data owners to identify risks in the Organisation's operations, match those risks to policies and procedures (including those developed as part of Specified Section 1 and Specified Action 2), and give effect to those policies and procedures.

- Delivery of training on data release policies and procedures must commence no later than 1 April 2020.
- Delivery of training on data governance policies and procedures must commence no later than 1 November 2020.
- The Organisation must provide OVIC with training course materials and a schedule for training when training commences.

Specified Action 4 - Reporting: The Organisation will provide OVIC with a report of its progress and compliance with Specified Action 1, Specified Action 2 and Specified Action 3, on 2 March 2020, 1 September 2020 and a final report on 1 March 2021.

3. Enforcement of this compliance notice

3.1 The Organisation must comply with this compliance notice.

3.2 If the Organisation does not comply with this compliance notice, the penalty is:

3.2.1 600 penalty units, in the case of an individual; and

3.2.2 3000 penalty units, in the case of a body corporate.

3.3 If the Organisation considers that it is not reasonably possible to take the action specified in this compliance notice within the period specified, the Organisation may apply to my office before the period of time specified in the compliance notice expires to extend the period of time specified in this compliance notice.

4. Application for review

4.1 An individual or organisation whose interests are affected by my decision to serve this compliance notice may apply to the Victorian Civil and Administrative Tribunal for review of my decision.

Rachel Dixon
Privacy and Data Protection Deputy Commissioner
05 August 2019