ETC5512: Data Ethics and Privacy

Table of contents

Data Ethics and Privacy 1
Data 2
Quick Look 2
Identification risks 4
Data Dictionary 5

Data Ethics and Privacy

When you make data open you don't always know how people will use it. This means we need to stop and think practically:

- How are people likely to want to use this data set?
- What might people being curious about using this data?

We also need to consider aspects of data ethics and privacy:

- Are there any applications of our data that we may want to protect against?
- How can protect people's privacy?

Quick concept check

What methods have we seen so far for de-identifying data? *Hint: Think back to the census week! Think back to our wild caught data types in week 1.*

Data

You can download the data for today's class from the development version of the R package `ggdibbler`.

```
if(!require("pak"))
  installed.packages("pak")

if(!require("ggdibbler"))
  pak::pak("harriet-mason/ggdibbler")

library(ggdibbler)
data("walktober")

# ?walktober
```

This data contains daily step counts during the 2025 Walktober challenge for five teams of four people. The participants are staff and PhD students in the Department of Econometrics and Business Statistics.

Quick Look

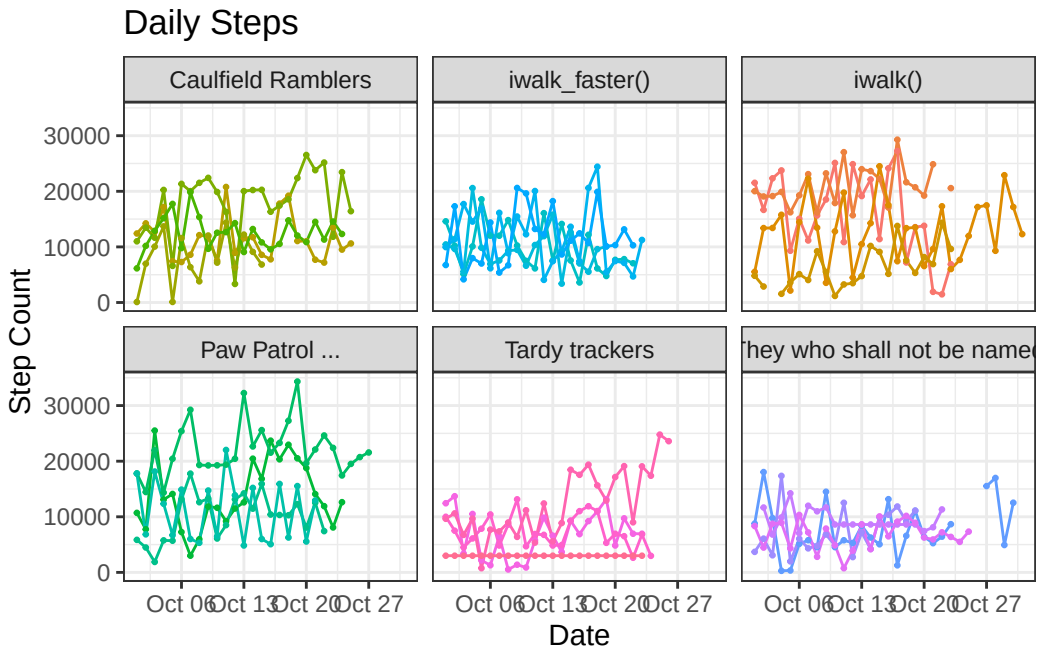
```
library(tidyverse)

glimpse(walktober)
```

```
Rows: 744
Columns: 4
$ team  <chr> "iwalk()", "iwalk()", "iwalk()", "iwalk()", "iwalk()", "iwalk()"~
$ name  <chr> "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "A",~
$ date  <chr> "01/10/25", "02/10/25", "03/10/25", "04/10/25", "05/10/25", "06/~
$ steps <int> 21526, 16656, 22329, 23753, 9307, 15062, 11155, 15660, 18526, 25~
```

```
walktober |>
  mutate(date = dmy(date)) |>
  ggplot(aes(x = date, y = steps, group = name, col = name)) +
  geom_line(size = 0.5) +
  geom_point(size = 0.5) +
  facet_wrap(~team) +
  theme_bw() +
```

```
  labs(title = "Daily Steps", y = "Step Count", x = "Date") +
  theme(legend.position = "none")
```




```
# walktober |>
#   mutate(date = dmy(date)) |>
#   group_by(team, date) |>
#   summarise(daily_steps = sum(steps, na.rm = TRUE)) |>
#   ungroup() |>
#   ggplot(aes(x = date, y = daily_steps, colour = team)) +
#   geom_line() +
#   geom_point() +
#   labs(title = "Daily Steps by Team",
#     x = NULL,
#     y = "Total Steps",
#     colour = "Team"
#   ) +
#   theme_bw()
```

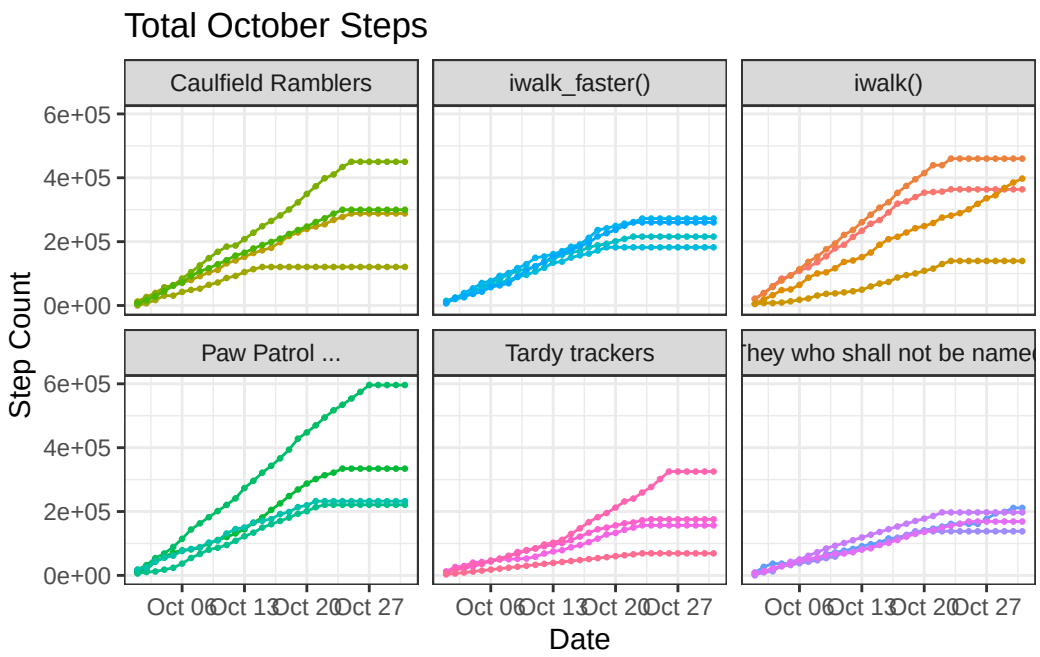
Question:

- What might you be curious about?
- What do you think others will be curious about?

This matters as when we de-identify data, we need to balance the data utility against preserving people's data privacy

Identification risks

```
walktober |>
  mutate(date = dmy(date)) |>
  group_by(team, name) |>
  arrange(date, .by_group = TRUE) |>
  mutate(total_steps = cumsum(replace_na(steps, 0))) |>
  ungroup() |>
  ggplot(aes(x = date, y = total_steps, group = name, col = name)) +
  geom_line(size = 0.5) +
  geom_point(size = 0.5) +
  facet_wrap(~team) +
  theme_bw() +
  labs(title = "Total October Steps", y = "Step Count", x = "Date") +
  theme(legend.position = "none")
```



Question:

- What do you think poses identification risks or re-identification risks in this data?

- Alternatively, if I told you that Kris, Maliny and I were in this data and you could ask for three additional variables to figure out who we were - what would you ask?

Data Dictionary

A data dictionary is different to meta data!

Metadata tells you about a dataset; a data dictionary tells you how to interpret the contents within it.
Data Dictionary – A structured document/catalog that defines the details of data(e.g., variables names, data types, allowed values, and descriptions). Think of it as the "user manual" for a dataset.
Metadata – Data about data — high-level information describing a dataset's context, such as who created it, when, its size, and format. Think of it as the "label on a tin can" that tells you what's inside without opening it.

Variable	Description	Type	Other Information
`team`	Original team name	`character string`	No fixed length, may contain emojis, has not been de-identified
`name`	References individual participants	`character string`	Individuals have been de-identified by replacing the names with a letter. Currently single letters only (<26 participants), but string length would increase if new participants added
`date`	Date of recorded step count	`character string`	Stored as character in raw data, should be converted to a date variable with AEST timezone for analysis
`steps`	Number of steps recorded by the participant on that date	`integer`	Sources include wearables, phone apps, and estimated workout proxies (biking, running, pilates). `NA` values occur due to non-reporting or reluctance to report small totals