Detecting Life Changes from Step Counts

July 18th, 2019

Erle Holgersen (@wuergh)

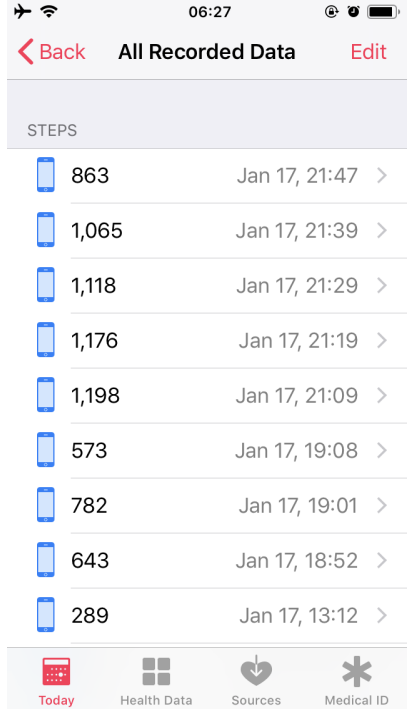
A Short Introduction to Me

- Statistician at Deep Genomics
  - R, Python
- Master's degree in statistics from Waterloo in 2017
- Interested in approaching statistical concepts through everyday problems
  - *Blog:* erle.io
- I've moved around a lot
  - *Past 5 years:* Ireland -> Canada -> UK -> Canada

Step Counts

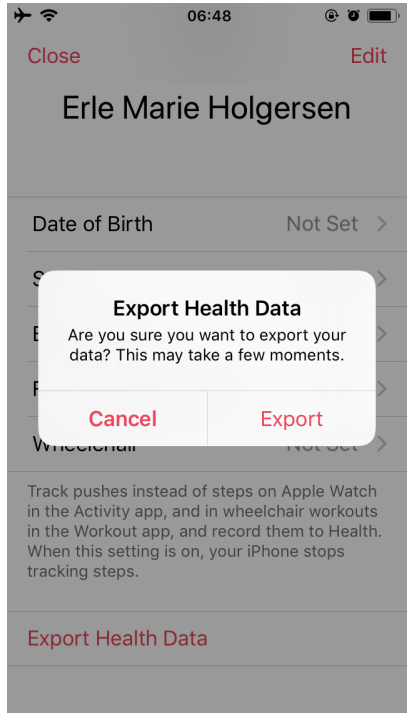
- Got my phone in March 2016
- Moves since then:

  Waterloo -> Toronto -> Waterloo -> Toronto -> London (UK) -> Toronto



Processing Data in R

- Export function in health app, email yourself the data



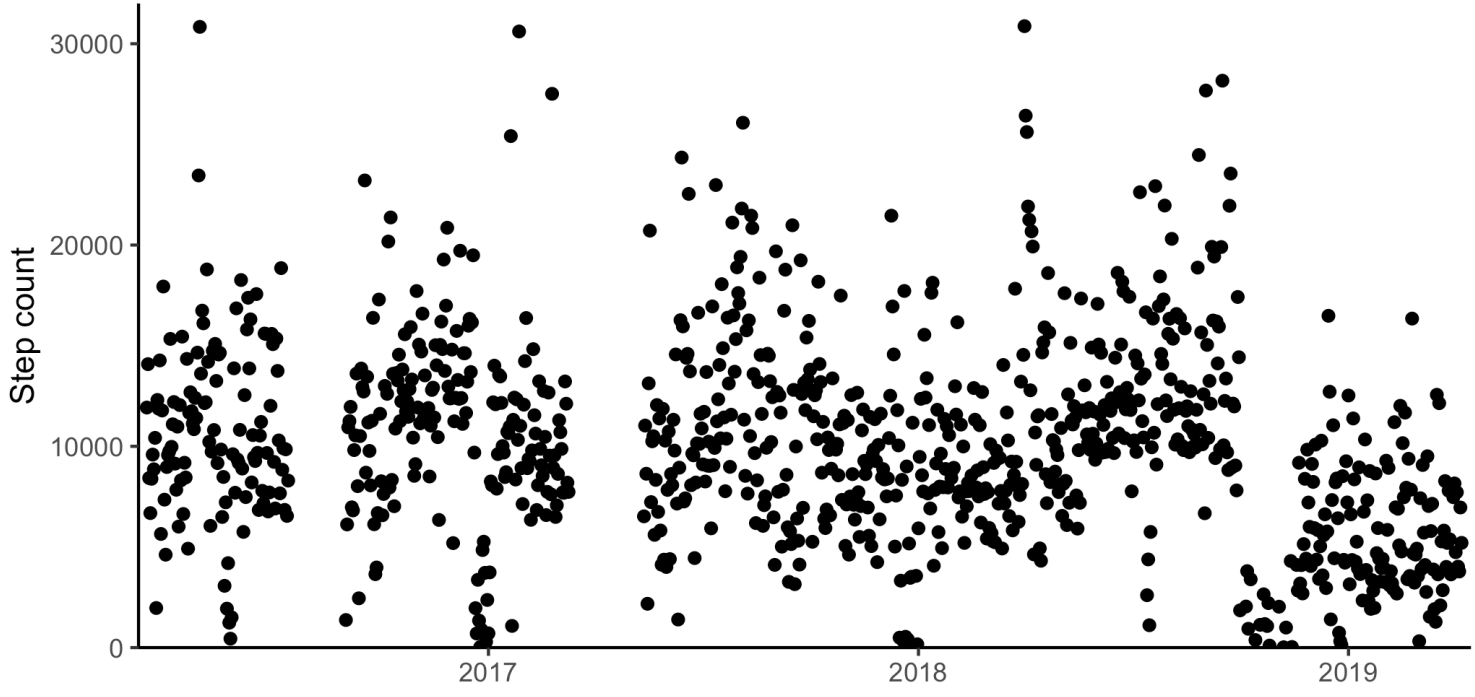
- Two XML files, interested in **export.xml**

```
<Record type="HKQuantityTypeIdentifierStepCount"
sourceName="Mycroft" sourceVersion="9.2.1"
device="&lt;&lt;HKDevice: 0x28057db80&gt;,
name:iPhone, manufacturer:Apple, model:iPhone,
hardware:iPhone6,1, software:9.2.1&gt;"
unit="count" creationDate="2016-03-17 19:54:24 -0400"
startDate="2016-03-17 19:42:57 -0400"
endDate="2016-03-17 19:44:19 -0400" value="117"/>
```

- XML package can be used to read XML files

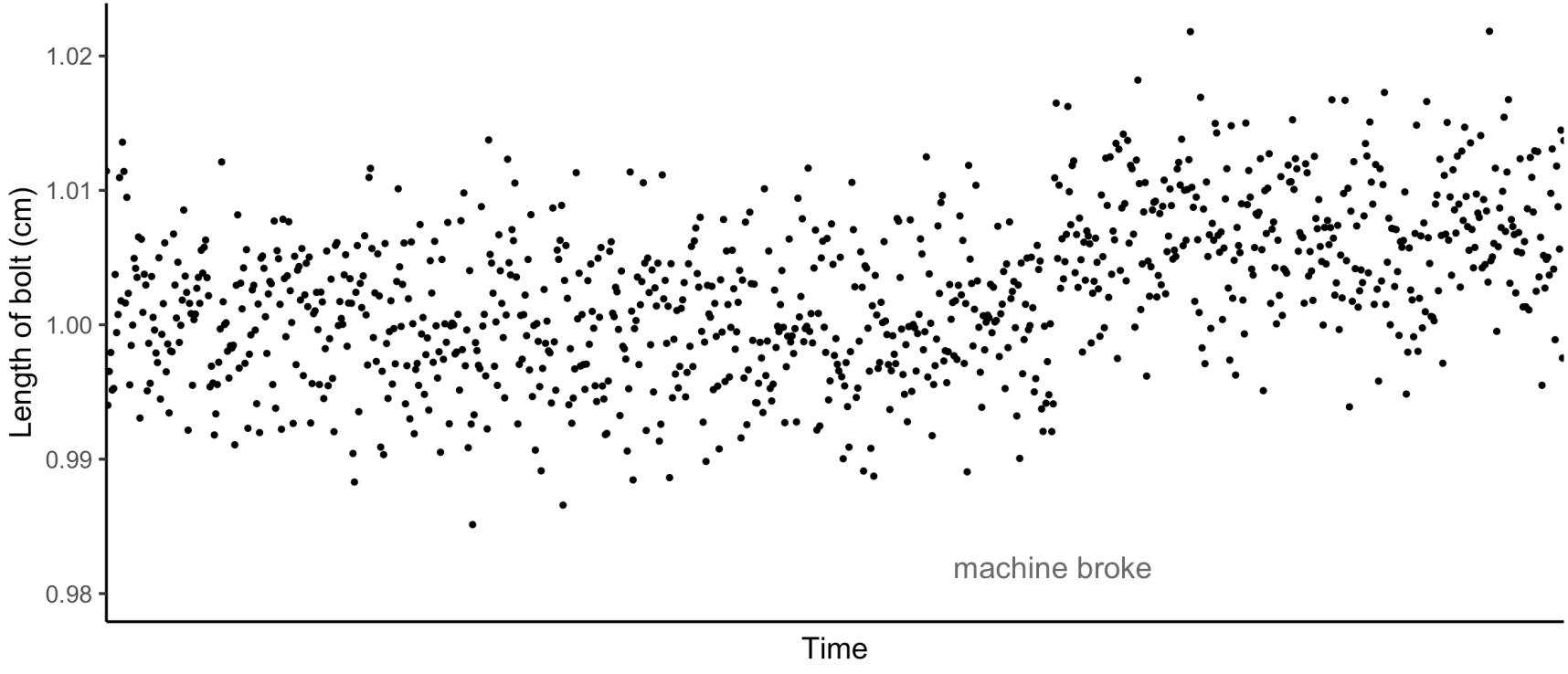
```
library(XML)

raw_data <- xmlParse('data/export.xml')
raw_data <- xmlToList(raw_data)
```

- Need to do some post-processing to parse the timestamps and convert to data frames
- Aggregated data to get the total number of steps per day


Step count
30000
20000
10000
0
2017
2018
2019

Changepoint Detection

- Method for identifying changes in a data generating process
- Often used for quality control in manufacturing



- Have a series of ordered observations $X_1, X_2, ..., X_n$
- Assume we can group observations into blocks so that within each block the observations follow the same distribution
    - *Machine working:* $X_1, X_2, ..., X_{650}$
    - *Machine broken:* $X_{651}, ..., X_{1000}$
    - Timestep 651 is known as a changepoint
- Choice in how to model data within each block
    - *Parametric:* Assume observations follow a specific distribution
        - e.g. normal distribution with changes in mean
    - *Non-parametric:* Don't assume anything about distribution

Nonparametric Changepoint Detection

- Zhu et al. 2014: *Nonparametric Maximum Likelihood Approach To Multiple Change-Point Problems*
- Don't assume any distribution, use binomial

$$X_1, ..., X_N \sim F_0(x)$$

$$\#\{X_i \leq u\} \sim \text{Binomial}(N, F_0(u))$$

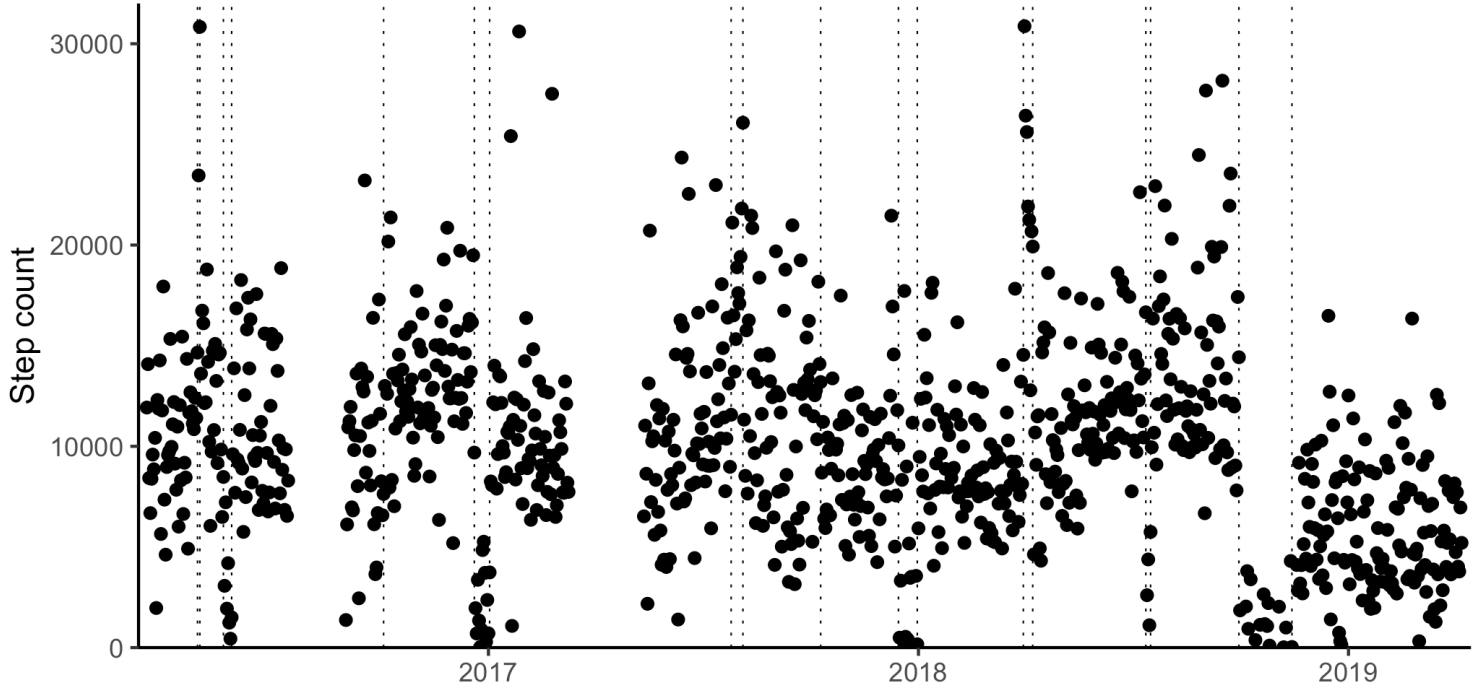
- R package: changepoint.np

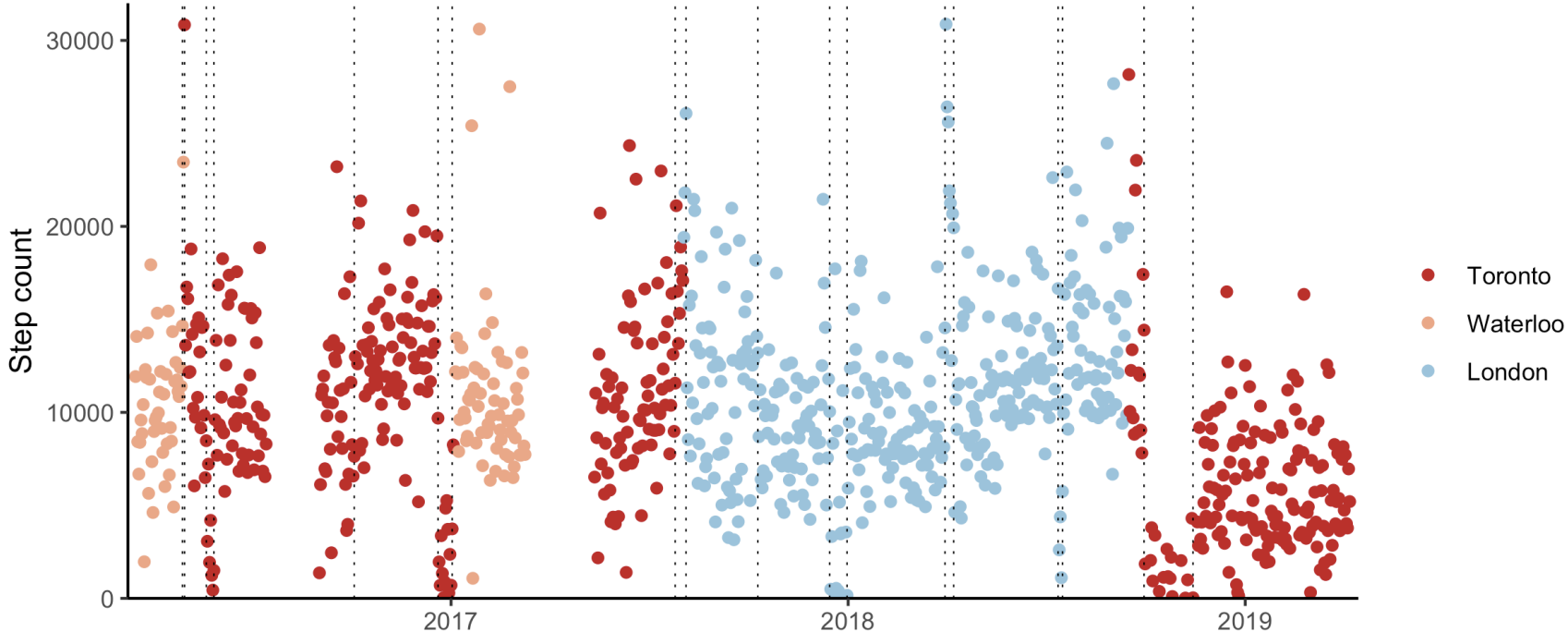
```
devtools::install.github("KayleaHaynes/changepoint.np")
```

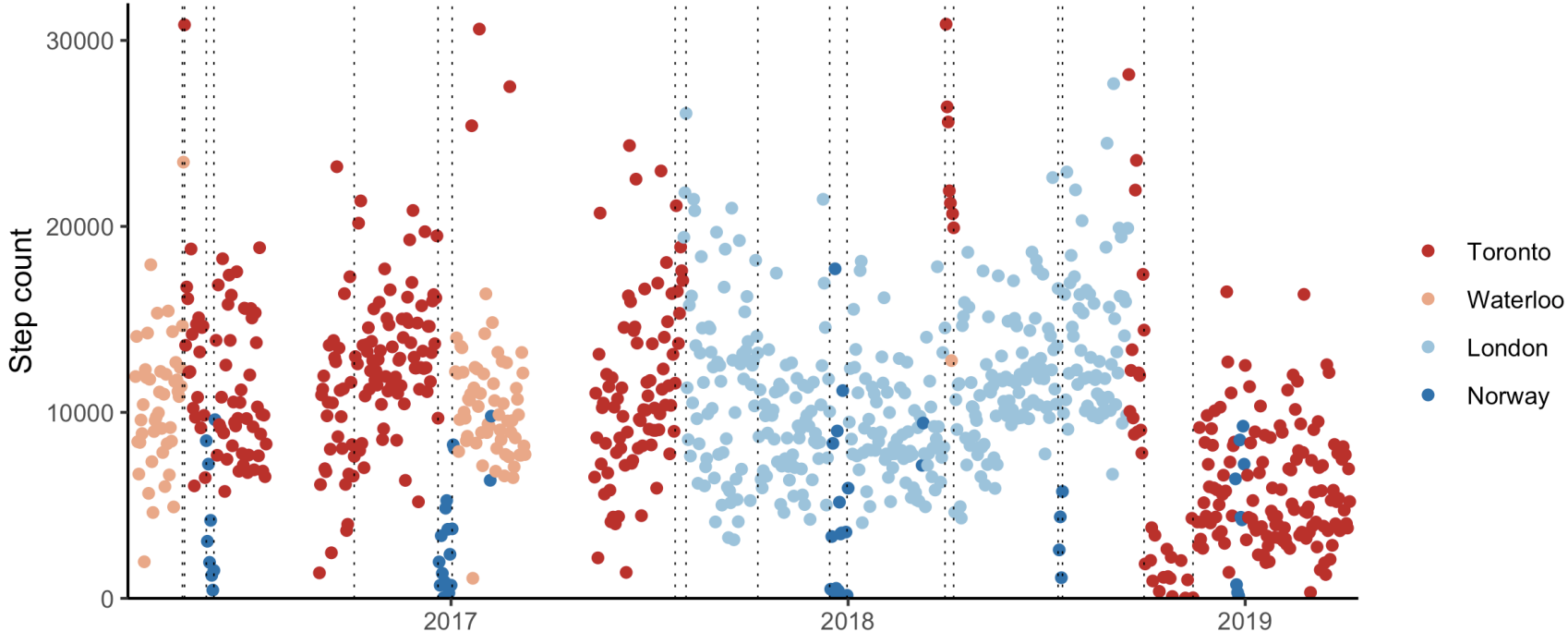
- `cpt.np` function calculates changepoints
- Optimal number of changepoints chosen based on a penalized likelihood criterion

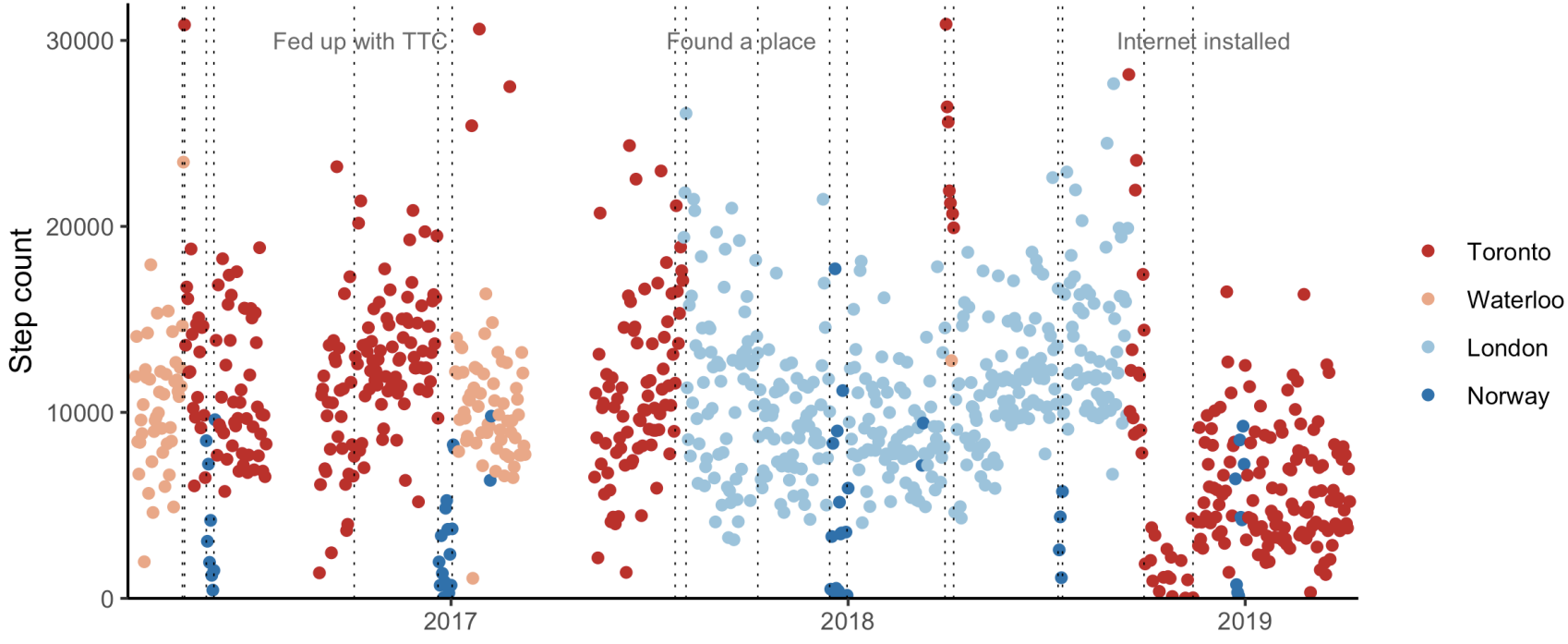
```
library(changepoint.np)

steps_cp <- cpt.np(steps$value)
change_indices <- attr(steps_cp, 'cpts')
```

- 18 changepoints in my step counts


Step count
30000
20000
10000
0
2017
2018
2019


Step count
30000
20000
10000
0
2017
2018
2019
Toronto
Waterloo
London


Step count
30000
20000
10000
0
2017
2018
2019
Toronto
Waterloo
London
Norway


Step count
30000
20000
10000
0
Fed up with TTC
Found a place
Internet installed
2017
2018
2019
Toronto
Waterloo
London
Norway

Conclusion

- Daily step count alone can identify life changes
- Is this useful?
  - I know where I've lived
  - Location data better for identifying house moves
  - Might still pick up other types of life changes