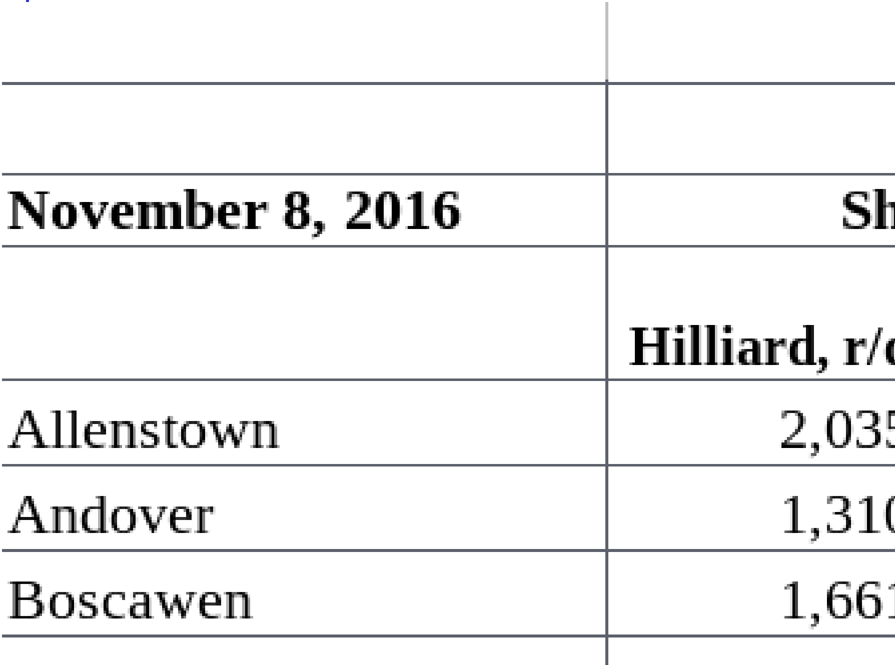
Extracting election data from spreadsheets

James Dunham, MIT MEDSL

July 3, 2018

The problem



November 8, 2016	Sh
	Hilliard, r/
Allenstown	2,035
Andover	1,310
Boscawen	1,661

Pretend it's tabular?

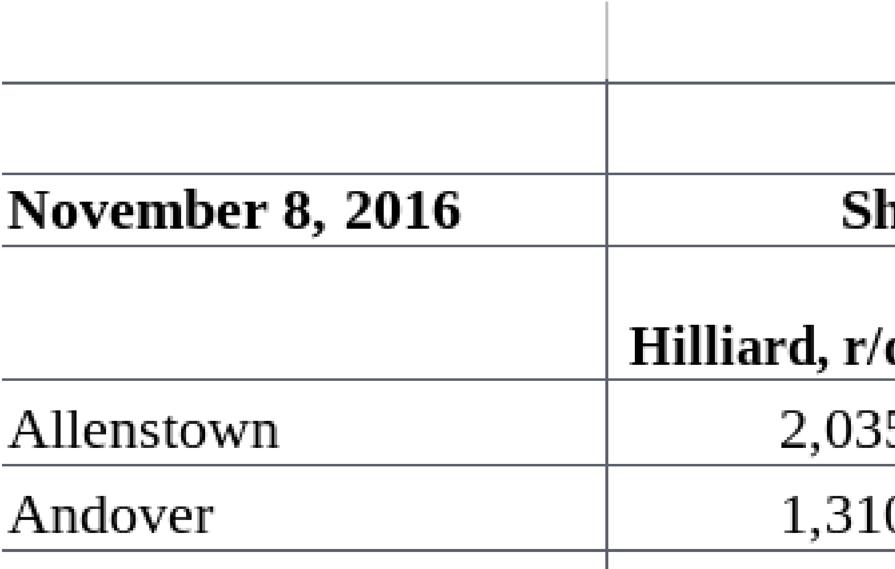
```r
library(dplyr)
library(stringr)
library(readxl)
library(medslcleaner)

merrimack_path = spreadsheet_example('merrimack')
read_excel(merrimack_path, col_names = FALSE) %>%
  head()
```

```
## # A tibble: 6 x 8
##   ..1        ..2               ..3    ..4     ..5    ..6
##   <chr>      <chr>             <chr>  <chr>   <chr>  <ch
## 1 <NA>       State of New Ham~ <NA>   <NA>    <NA>   <NA
## 2 <NA>       Merrimack County~ <NA>   <NA>    <NA>   <NA
## 3 42682      Sheriff           <NA>   Attorn~ <NA>   Tre
## 4 <NA>       Hilliard, r/d     Scatt~ Murray~ Scatt~ Han
## 5 Allenstown 2035              8      2013    3      129
## 6 Andover    1310              3      1277    <NA>   714
```

Alternative solution

1. Identify which cells are **data** and which are **headers**
2. Define the **relationships** between data cells and header cells



November 8, 2016	**Sh**
	Hilliard, r/
Allenstown	2,035
Andover	1,31

Tools

R packages:

- `tidyverse`
- `tidyxl`
- `unpivotr`
- `medslcleaner`

Read the data

```
library(medslcleaner)
library(tidyverse)
library(tidyxl)
library(unpivotr)

# For this example only: get path to the spreadsheet
merrimack_path = spreadsheet_example('merrimack')

d = xlsx_cells(merrimack_path, sheet = 1)  # from tidyxl
```

Representation in R

```
d %>%
  select(address, row, col, data_type, character, numeric)
  head()
```

```
## # A tibble: 6 x 6
##   address   row   col data_type character
##   <chr>   <int> <int> <chr>     <chr>
## 1 A1          1     1 blank     <NA>
## 2 B1          1     2 character State of New Hampshire -
## 3 C1          1     3 blank     <NA>
## 4 D1          1     4 blank     <NA>
## 5 E1          1     5 blank     <NA>
## 6 F1          1     6 blank     <NA>
```

Associate headers with cells

```
d = d %>%
  filter(row > 2) %>%
  behead('W', 'precinct')

d %>%
  filter(row > 4) %>%
  arrange(row, col) %>%
  select(address, row, col, character, numeric, precinct)
```

```
## # A tibble: 35 x 6
##    address   row   col character numeric precinct
##    <chr>   <int> <int> <chr>         <dbl> <chr>
##  1 B5          5     2 <NA>           2035 Allenstown
##  2 C5          5     3 <NA>              8 Allenstown
##  3 D5          5     4 <NA>           2013 Allenstown
##  4 E5          5     5 <NA>              3 Allenstown
##  5 F5          5     6 <NA>           1296 Allenstown
##  6 G5          5     7 <NA>            699 Allenstown
##  7 H5          5     8 <NA>              2 Allenstown
```

Associate headers with cells

```
d = d %>%
  behead('NNW', 'office') %>%
  behead('N', 'candidate')

d %>%
  arrange(row, col) %>%
  select(address, row, col, character, numeric, office, can
```

```
## # A tibble: 35 x 7
##    address   row   col character numeric office     candi
##    <chr>   <int> <int> <chr>       <dbl> <chr>      <chr>
##  1 B5          5     2 <NA>         2035 Sheriff    Hilli
##  2 C5          5     3 <NA>            8 Sheriff    Scatt
##  3 D5          5     4 <NA>         2013 Attorney   Murra
##  4 E5          5     5 <NA>            3 Attorney   Scatt
##  5 F5          5     6 <NA>         1296 Treasurer " Han
##  6 G5          5     7 <NA>          699 Treasurer Rodri
##  7 H5          5     8 <NA>            2 Treasurer Scatt
##  8 B6          6     2 <NA>         1310 Sheriff    Hilli
```

Resources

- Spreadsheet Munging Strategies: https://nacnudus.github.io/spreadsheet-munging-strategies