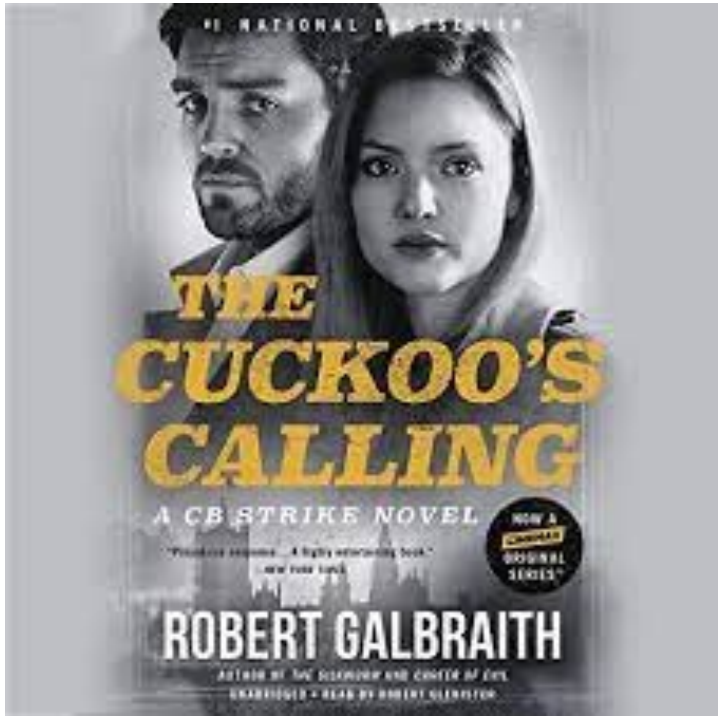
# text analytics

# text mining applications

- spam filters for email
- document relevancy in search engines
- summarization and trend analysis of social media
- automated grading of student essays
- author attribution (who wrote Shakespeare plays?)
- AI written news stories

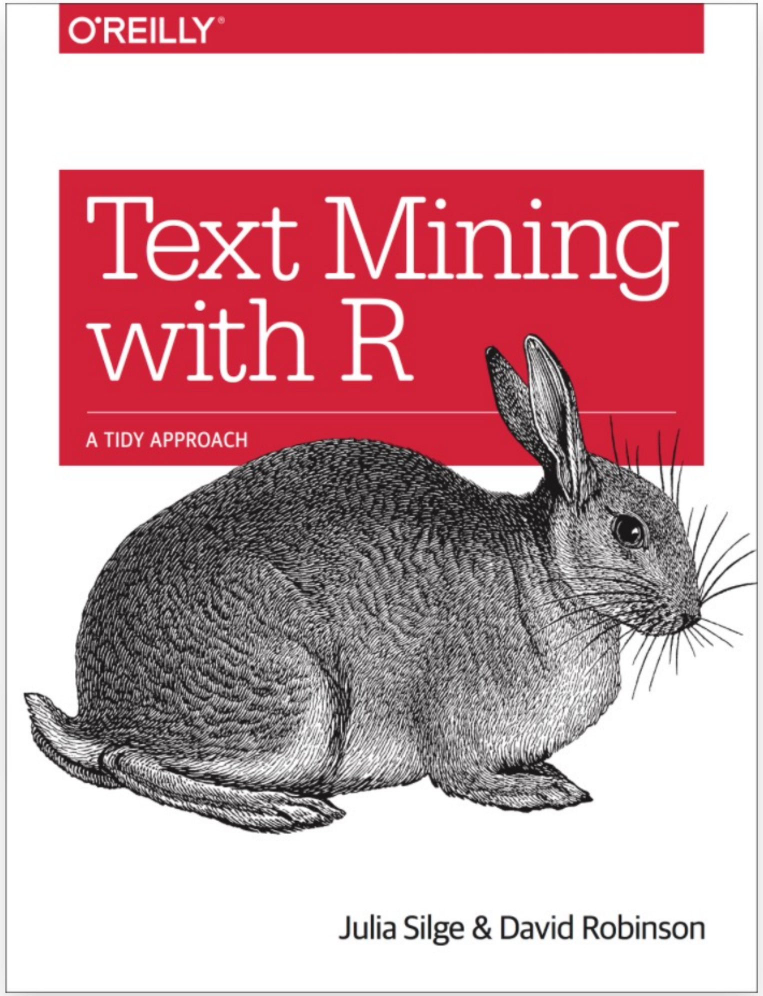
# cuckoo’s calling analysis

- Patrick Juola (Duquesne University)
- JGAAP (Java Graphical Authorship Attribution Program)
- Distribution of word lengths
- 100 most common words
- Distribution of 4-grams (4 consecutive letters)
- Distribution of bi-grams



# Text mining w/ R

https://www.tidytextmining.com/index.html



# Tidy text format

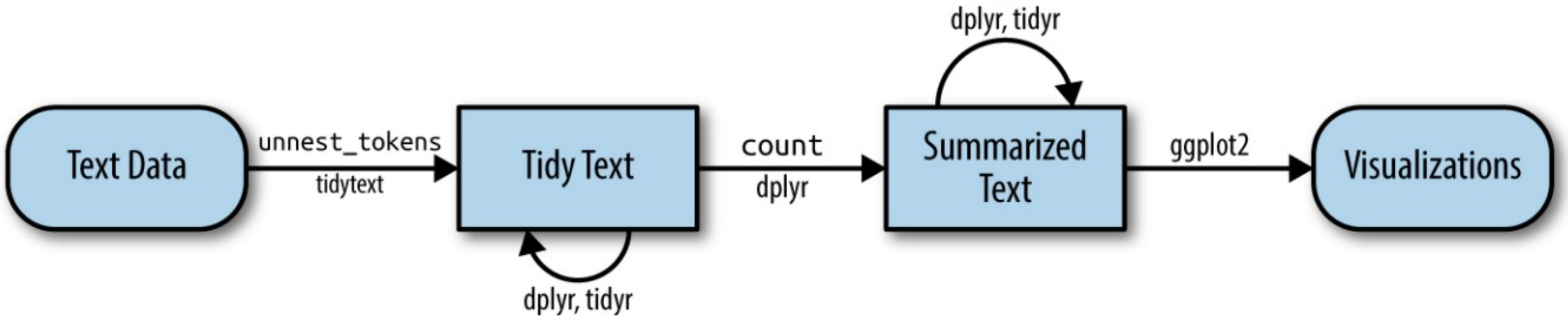
```
text <- c("Because I could not stop for Death -", "He kindly
stopped for me -", "The Carriage held but just Ourselves -",
"and Immortality")
```

```
library(dplyr)
text_df <- tibble(line = 1:4, text = text)

text_df
#> # A tibble: 4 x 2
#>    line text
#>   <int> <chr>
#> 1     1 Because I could not stop for Death -
#> 2     2 He kindly stopped for me -
#> 3     3 The Carriage held but just Ourselves -
#> 4     4 and Immortality
```

# Tokens

```
library(tidytext)

text_df %>%
  unnest_tokens(word, text)
#> # A tibble: 20 x 2
#>     line word
#>    <int> <chr>
#>  1     1 because
#>  2     1 i
#>  3     1 could
#>  4     1 not
#>  5     1 stop
#>  6     1 for
#>  7     1 death
#>  8     2 he
#>  9     2 kindly
#> 10     2 stopped
#> # … with 10 more rows
```

# Workflow


Text Data
unnest_tokens
tidytext
Tidy Text
dplyr, tidyr
count
dplyr
dplyr, tidyr
Summarized Text
ggplot2
Visualizations

# Stopwords

```
 tidy_books %>%
   count(word, sort = TRUE)
# A tibble: 14,520 x 2
  word      n
  <chr> <int>
 1 the   26351
 2 to    24044
 3 and   22515
 4 of    21178
 5 a     13408
 6 her   13055
 7 i     12006
 8 in    11217
 9 was   11204
10 it    10234
```

```
tidy_books %>%
  anti_join(stop_words) %>%
  count(word, sort = TRUE)
#> # A tibble: 13,914 x 2
#>    word       n
#>    <chr>  <int>
#>  1 miss    1855
#>  2 time    1337
#>  3 fanny    862
#>  4 dear     822
#>  5 lady     817
#>  6 sir      806
#>  7 day      797
#>  8 emma     787
#>  9 sister   727
#> 10 house    699
#> # … with 13,904 more rows
```

# Sentiment datasets

•AFINN from Finn Årup Nielsen,
•bing from Bing Liu and collaborators,
and
•nrc from Saif Mohammad and Peter
Turney.

# Example

```
library(tidytext)

get_sentiments("afinn")
```

```
#> # A tibble: 2,477 x 2
#>    word        value
#>    <chr>       <dbl>
#>  1 abandon        -2
#>  2 abandoned      -2
#>  3 abandons       -2
#>  4 abducted       -2
#>  5 abduction      -2
#>  6 abductions     -2
#>  7 abhor          -3
#>  8 abhorred       -3
#>  9 abhorrent      -3
#> 10 abhors         -3
#> # … with 2,467 more rows
```

# Example

```
text=c("I hate the dentist","I love candy")
text_df <- tibble(line = 1:2, text = text)


text_df %>%
  unnest_tokens(word, text) %>%
  inner_join(sentiment_table,by="word") %>%
  group_by(line) %>%
  summarise(avg_sentiment=mean(value))
```

```
# A tibble: 2 x 2
   line avg_sentiment
* <int>         <dbl>
1     1            -3
2     2             3
```