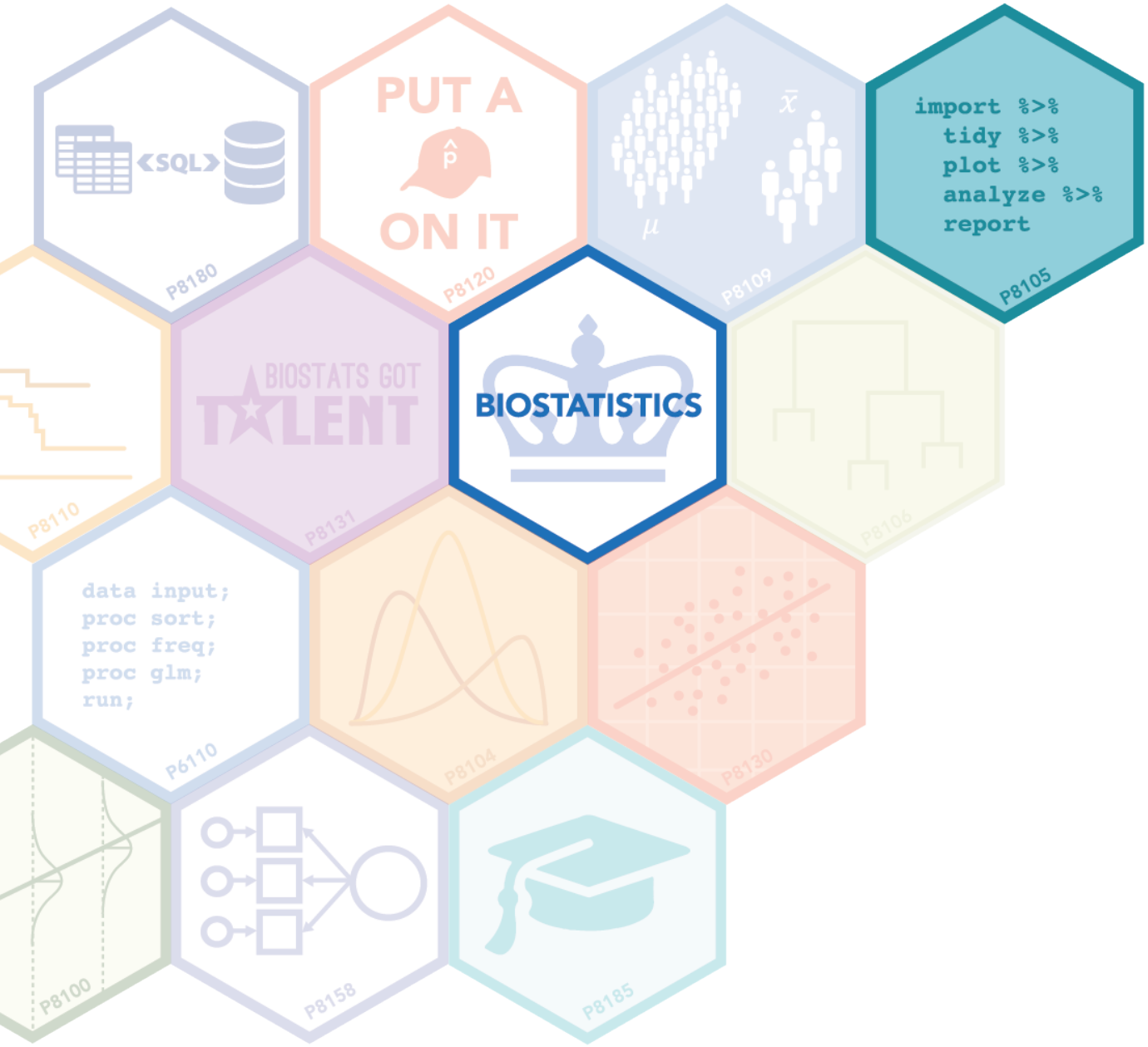


TIDY TEXT

Jeff Goldsmith, PhD
Department of Biostatistics

Text data

- Written information
  - Sentences
  - Tweets
  - Descriptions
  - Books

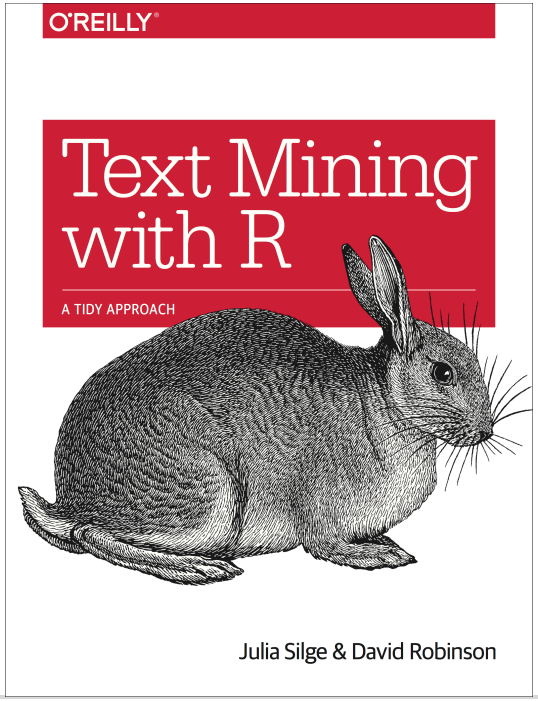
- Stored as strings
- Made up of “tokens”, which is a meaningful unit of text
  - Words; sentences; paragraphs; etc

Tidy text

- Need to organize text data around tokens
  - If your data contain whole tweets as a variable and your tokens are words, your data aren't "tidy"
  - "Un-nesting" is a common step

- Once you have tidy text data, you need to analyze it

- The `tidytext` package contains useful tools



Words

- Stop words are common but don't contain information
  - "the", "of", "and", etc.
- `tidytext` has a dataset of stop words called `stop_words`
- Remove these from your tidy text data using an anti-join

- Word frequency is often very informative
  - Count words in tidy text datasets using `group_by` and `summarize`

Relative frequencies

- Comparisons across groups are often informative
- Word counts alone may be misleading – group sizes may differ
- If only there were a way to see if words were more likely to appear in one group than in another group …

- Odds ratios! Yay!
- We'll use an approximate odds ratio, which guards against division-by-zero for uncommon words

$$\log \text{OR} \approx \ln \frac{\left[\frac{\text{word count } +1}{\text{total count } +1} \mid \text{Comparison}\right]}{\left[\frac{\text{word count } +1}{\text{total count } +1} \mid \text{Reference}\right]}$$

Sentiments

- Words convey sentiments
  - "Happy" is a happy word :-)
  - "Sad" is a sad word :-(

- Lexicons can map words to the sentiments they convey
  - `tidytext` contains several sentiment lexicons
  - Join to a tidy text dataset using joins
  - Construct overall score for a sentence / phrase by aggregating across individual words