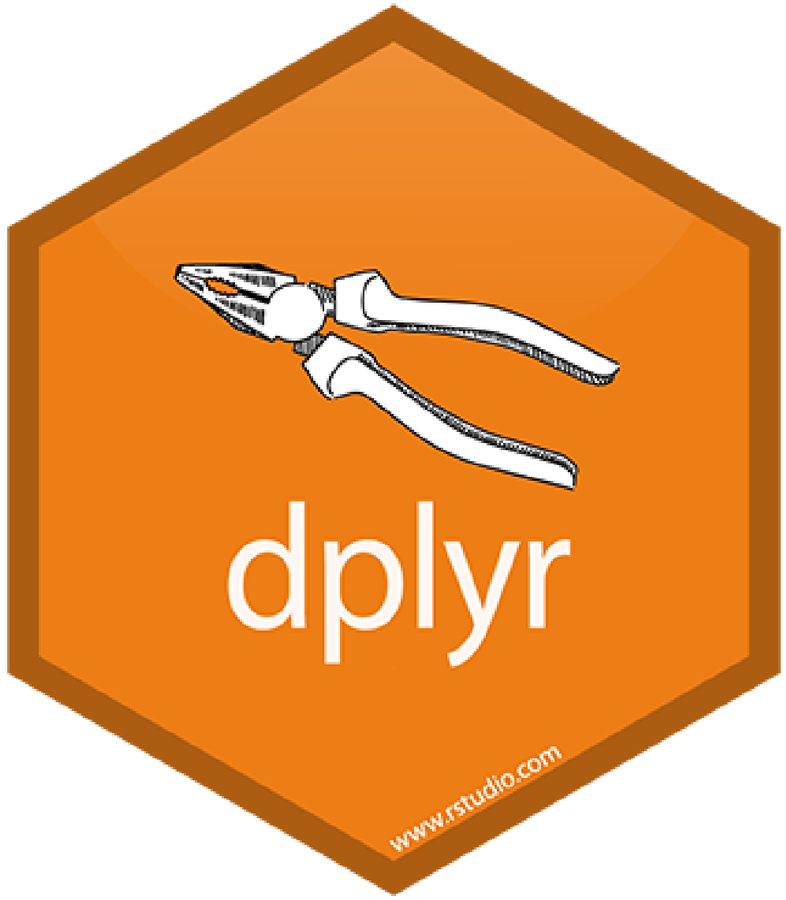
Text as data

INTRODUCTION TO TEXT ANALYSIS IN R

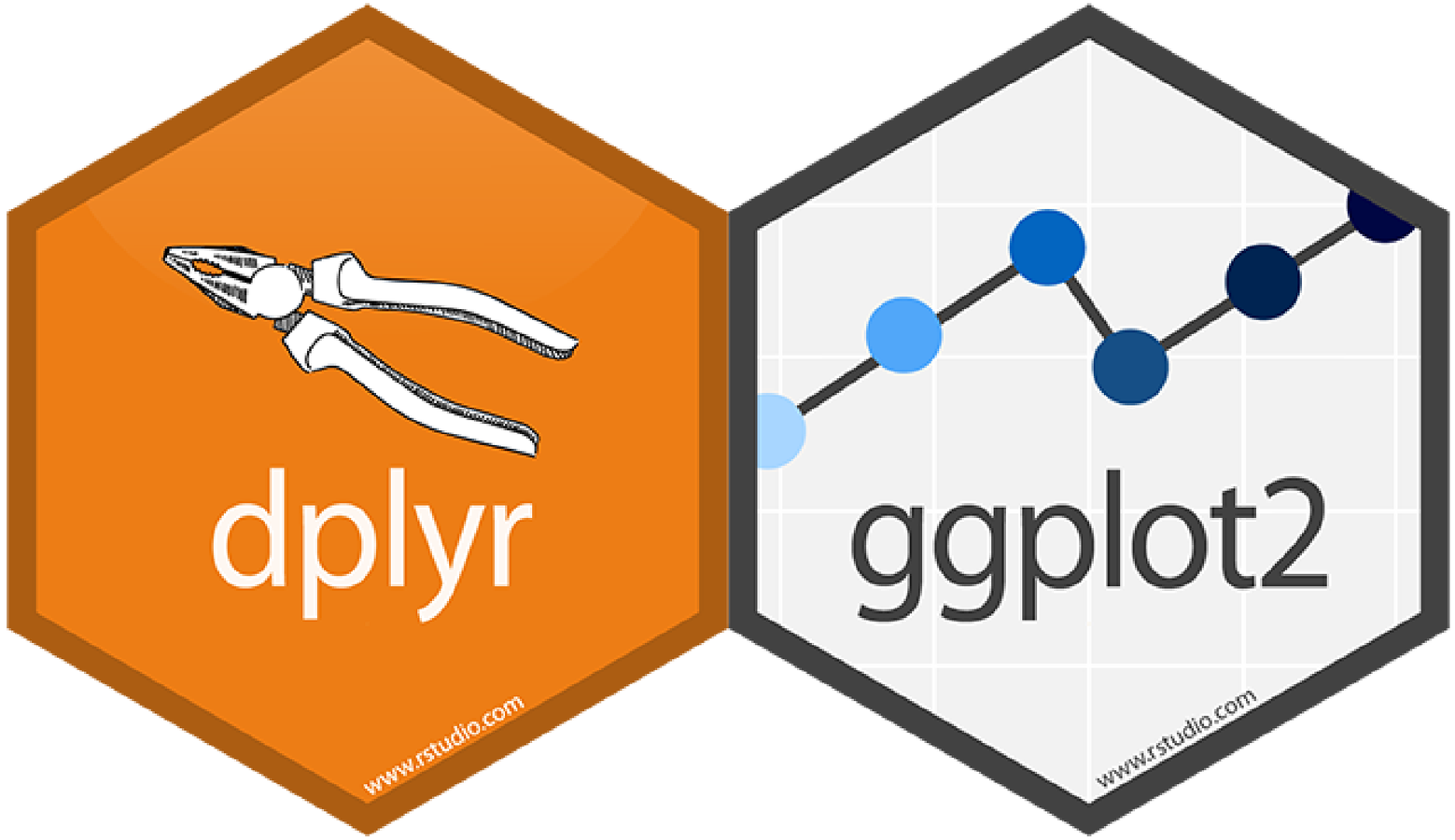
Maham Faisal Khan
Senior Data Science Content Developer, DataCamp

datacamp

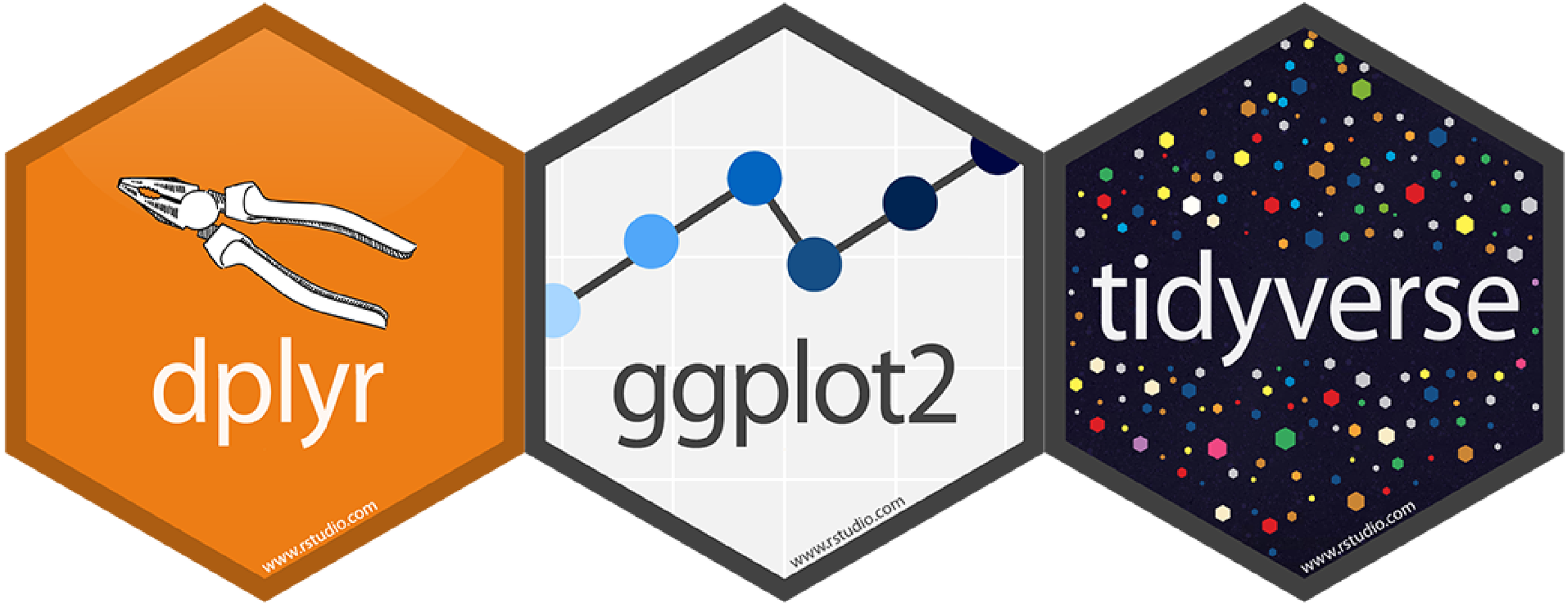
Using the tidyverse



Using the tidyverse



Using the tidyverse



Loading packages

```
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.2.1 --
v ggplot2 3.0.0     v purrr   0.2.5
v tibble  2.0.0     v dplyr   0.7.8
v tidyr   0.8.2     v stringr 1.3.1
v readr   1.1.1     v forcats 0.3.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

Importing review data

```
review_data <- read_csv("Roomba Reviews.csv")
review_data
```

```
# A tibble: 1,833 x 4
   Date     Product                  Stars Review
   <chr>    <chr>                    <dbl> <chr>
 1 2/28/15  iRobot Roomba 650 fo…        5 You would not believe how well...
 2 1/12/15  iRobot Roomba 650 fo…        4 You just walk away and it does...
 3 12/26/13 iRobot Roomba 650 fo…        5 You have to Roomba proof your...
 4 8/4/13   iRobot Roomba 650 fo…        3 Yes, its a fascinating, albeit...
# … with 1,829 more rows
```

Using filter() and summarize()

```
review_data %>%
  filter(product == "iRobot Roomba 650 for Pets") %>%
  summarize(stars_mean = mean(stars))
```

```
# A tibble: 1 x 1
  stars_mean
       <dbl>
1       4.49
```

Using group_by() and summarize()

```
review_data %>%
  group_by(product) %>%
  summarize(stars_mean = mean(stars))
```

```
# A tibble: 2 x 2
  product                                  stars_mean
  <chr>                                         <dbl>
1 iRobot Roomba 650 for Pets                     4.49
2 iRobot Roomba 880 for Pets and Allergies       4.42
```

Unstructured data

```
review_data %>%
  group_by(product) %>%
  summarize(review_mean = mean(review))
```

```
Warning messages:
1: In mean.default(review) :
  argument is not numeric or logical: returning NA
2: In mean.default(review) :
  argument is not numeric or logical: returning NA
```

Let's practice!

INTRODUCTION TO TEXT ANALYSIS IN R

Counting categorical data

INTRODUCTION TO TEXT ANALYSIS IN R

Maham Faisal Khan
Senior Data Science Content Developer

Column types

```
review_data
```

```
# A tibble: 1,833 x 4
   date     product                       stars review
   <chr>    <chr>                         <dbl> <chr>
 1 2/28/15  iRobot Roomba 650 fo…             5 You would not believe how well...
 2 1/12/15  iRobot Roomba 650 fo…             4 You just walk away and it does...
 3 12/26/13 iRobot Roomba 650 fo…             5 You have to Roomba proof your...
 4 8/4/13   iRobot Roomba 650 fo…             3 Yes, its a fascinating, albeit...
 5 12/22/15 iRobot Roomba 650 fo…             5 Years ago I bought one of the...
# … with 1,828 more rows
```

Summarizing with n()

```
review_data %>%
  summarize(number_rows = n())
```

```
# A tibble: 1 x 1
  number_rows
        <int>
1        1833
```

Summarizing with n()

```
review_data %>%
  group_by(product) %>%
  summarize(number_rows = n())
```

```
# A tibble: 2 x 2
  product                                  number_rows
  <chr>                                          <int>
1 iRobot Roomba 650 for Pets                       633
2 iRobot Roomba 880 for Pets and Allergies        1200
```

Summarizing with count()

```
review_data %>%
  count(product)
```

```
# A tibble: 2 x 2
  product                                      n
  <chr>                                    <int>
1 iRobot Roomba 650 for Pets                 633
2 iRobot Roomba 880 for Pets and Allergies  1200
```

Summarizing with count()

```
review_data %>%
  count(product) %>%
  arrange(desc(n))
```

```
# A tibble: 2 x 2
  product                                          n
  <chr>                                        <int>
1 iRobot Roomba 880 for Pets and Allergies  1200
2 iRobot Roomba 650 for Pets                 633
```

Let's practice!

INTRODUCTION TO TEXT ANALYSIS IN R

Tokenizing and cleaning

INTRODUCTION TO TEXT ANALYSIS IN R

Maham Faisal Khan
Senior Data Science Content Developer

Using tidytext

Tokenizing text

Some natural language processing (NLP) vocabulary:

- Bag of words: Words in a document are independent
- Every separate body of text is a document
- Every unique word is a term
- Every occurrence of a term is a token
- Creating a bag of words is called tokenizing

Using unnest_tokens()

```
tidy_review <- review_data %>%
  unnest_tokens(word, review)
tidy_review
```

```
# A tibble: 229,481 x 4
   date    product                          stars word
   <chr>   <chr>                            <dbl> <chr>
 1 2/28/15 iRobot Roomba 650 for Pets           5 you
 2 2/28/15 iRobot Roomba 650 for Pets           5 would
 3 2/28/15 iRobot Roomba 650 for Pets           5 not
# … with 229,478 more rows
```

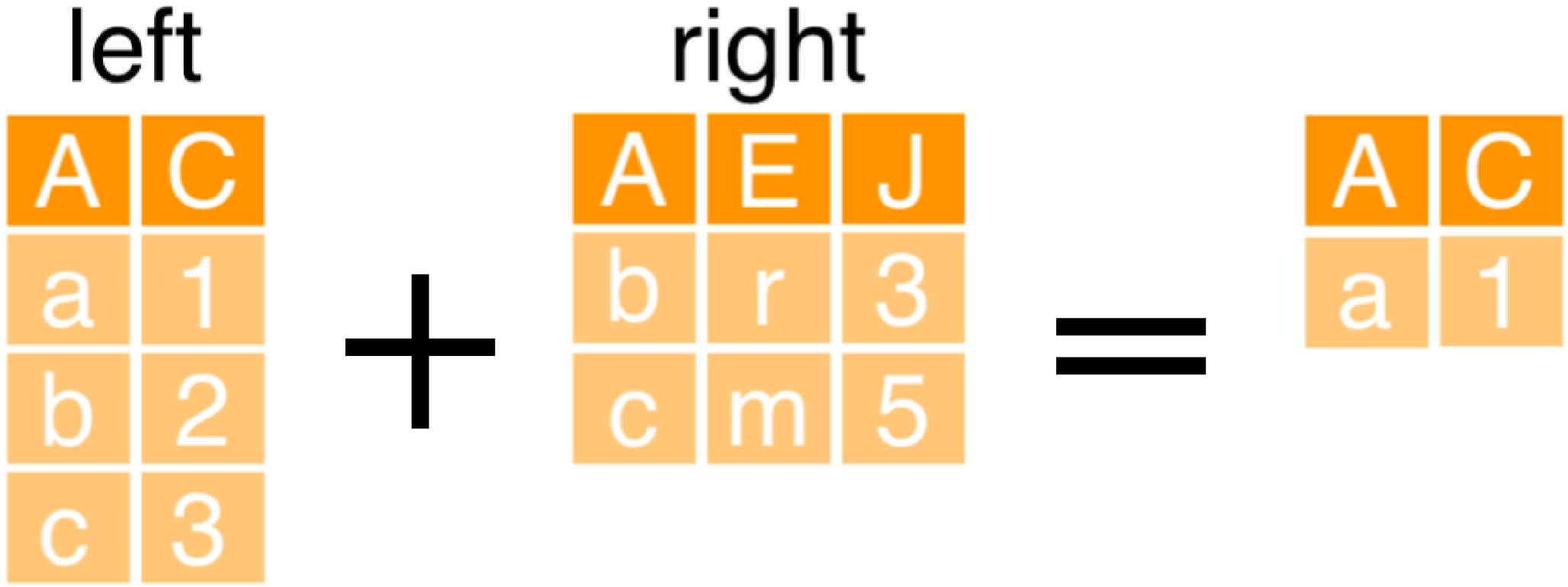
Counting words

```
tidy_review %>%
  count(word) %>%
  arrange(desc(n))
```

```
# A tibble: 10,310 x 2
   word      n
   <chr> <int>
 1 the   11785
 2 it     7905
 3 and    6794
# … with 10,307 more rows
```

Using anti_join()

- We'd like to remove stop words from our tidied data frame
- We'll use joins to do this



Using anti_join()

```
tidy_review2 <- review_data %>%
  unnest_tokens(word, review) %>%
  anti_join(stop_words)
tidy_review2
```

```
# A tibble: 78,868 x 4
   date     product                       stars word
   <chr>    <chr>                         <dbl> <chr>
 1 1/12/15  iRobot Roomba 650 for Pets        4 walk
 2 1/12/15  iRobot Roomba 650 for Pets        4 rest
# … with 78,866 more rows
```

Counting words again

```
tidy_review2 %>%
  count(word) %>%
  arrange(desc(n))
```

```
# A tibble: 9,672 x 2
   word          n
   <chr>     <int>
 1 roomba     2286
 2 clean      1204
 3 vacuum      989
# … with 9,669 more rows
```

Let's practice!

INTRODUCTION TO TEXT ANALYSIS IN R