Getting / cleaning data 2

Regular expressions

Regular expressions

We've already done some things to manipulate strings. For example, if we wanted to separate "Name" into last name and first name (including title), we could actually do that with the `separate` function:

```
titanic_train %>%
  select(Name) %>%
  slice(1:3) %>%
  separate(Name, c("last_name", "first_name"), sep = ", ")
```

```
##   last_name                                        first_name
## 1    Braund                                    Mr. Owen Harris
## 2   Cumings Mrs. John Bradley (Florence Briggs Thayer)
## 3 Heikkinen                                       Miss. Laina
```

Regular expressions

Notice that `separate` is looking for a regular pattern (“,”) and then doing something based on the location of that pattern in each string (splitting the string).

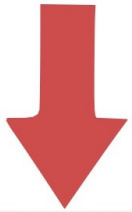
There are a variety of functions in R that can perform manipulations based on finding regular patterns in character strings.

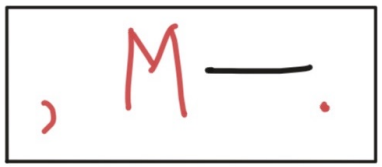
Regular expressions

Braund, Mr. Owen Harris

Cumings, Mrs. John Bradley (Florence Briggs Thayer)

Heikkinen, Miss. Laina





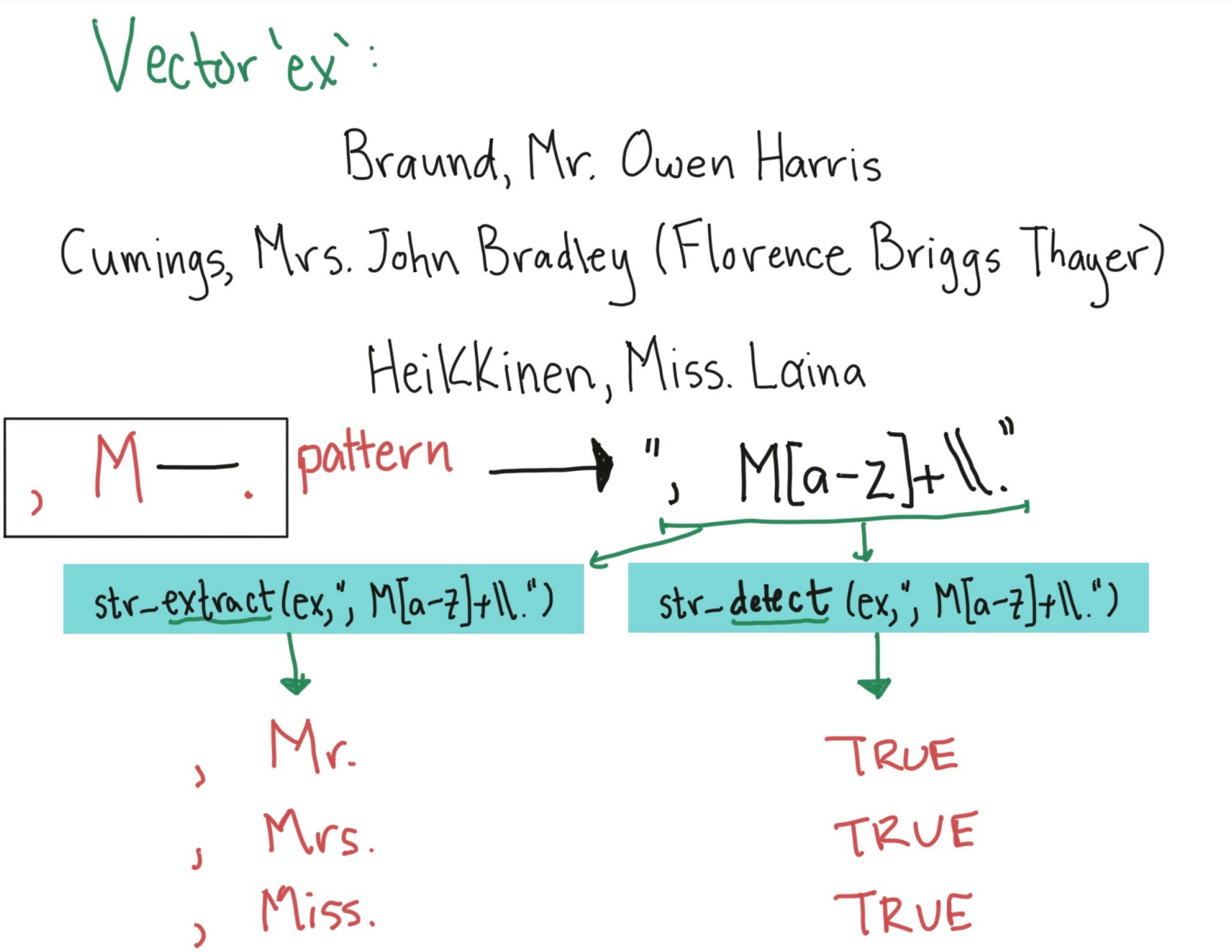
pattern

Braund, Mr. Owen Harris

Cumings, Mrs. John Bradley (Florence Briggs Thayer)

Heikkinen, Miss. Laina

Regular expressions



Regular expressions

pattern: "Mr"

strings	str_extract result	str_detect result
Mr.	Mr	TRUE
Mrs.	Mr	TRUE
Miss.	NA	FALSE
Dr.	NA	FALSE