

Scraping Twitter

Carolina A. de Lima Salge
Assistant Professor
Terry College of Business
University of Georgia

*Business Intelligence
Spring 2021*



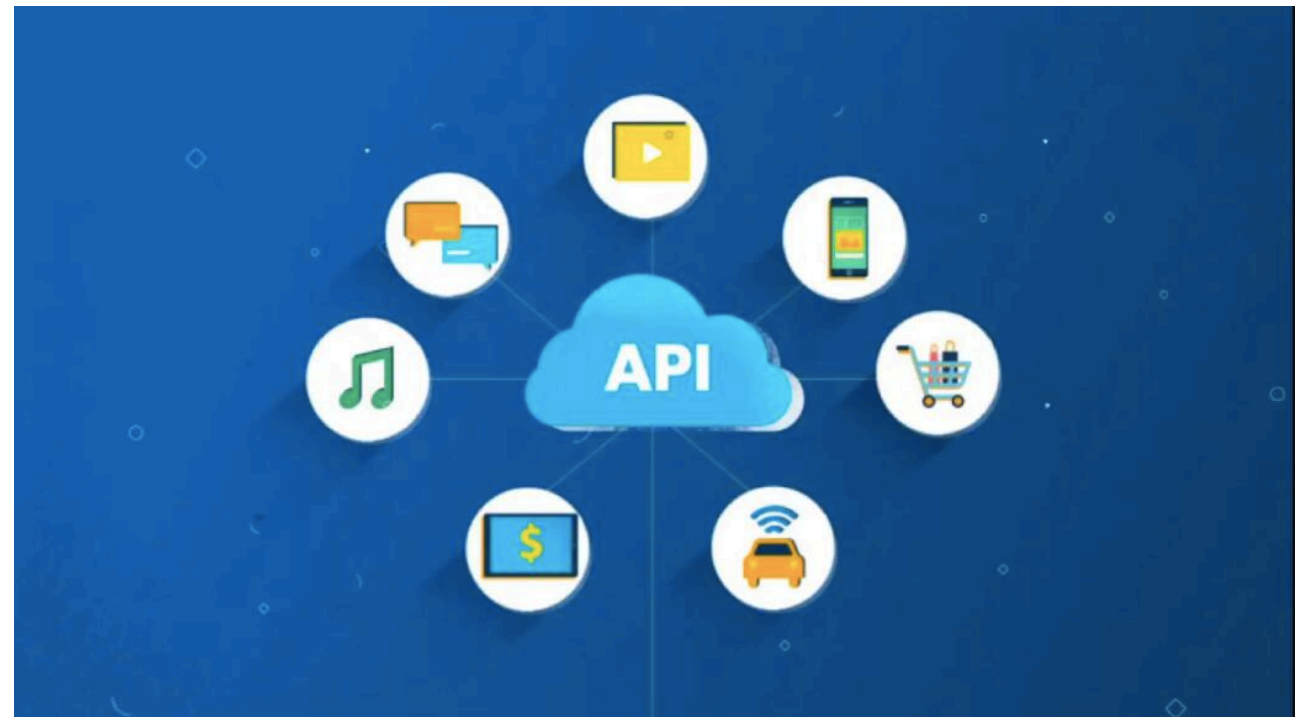
Terry College of Business
UNIVERSITY OF GEORGIA



```
1 install.packages("httr")
2
3 require("httr")
4
5 install.packages("jsonlite")
6
7 require("jsonlite")
8
9
10
11 (Top Level) ▾
```

Console ~ /Google Drive/Intrinio/Focus/Marketing/CUB Lecture/ ↗
/var/folders/15/2z483nk52b92jnsbcs58_15c0000gn/T//RtmpJBua8N/downloaded_packages
> require("httr")
> install.packages("jsonlite")
Error in install.packages : Updating loaded packages
> install.packages("jsonlite")
Installing package into '/Users/Andrew/Library/R/3.4/library'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.4/jsonlite_1.5.tgz'
Content type 'application/x-gzip' length 1114207 bytes (1.1 MB)
downloaded 1.1 MB

The downloaded binary packages are in
/var/folders/15/2z483nk52b92jnsbcs58_15c0000gn/T//RtmpJBua8N/downloaded_packages
> require("jsonlite")
> |



Web Scraping

Three common ways to scrape data

- Basic (HTML)
- APIs (Twitter, Yelp)
- Advanced (Javascript)
 - Rselenium: <https://rpubs.com/johndharrison/RSelenium-Basics>



Web Scraping

Three common ways to scrape data

- Basic (HTML)
- APIs (Twitter, Yelp)
- Advanced (Javascript)
 - Relenium: <https://rpubs.com/johndharrison/RSelenium-Basics>

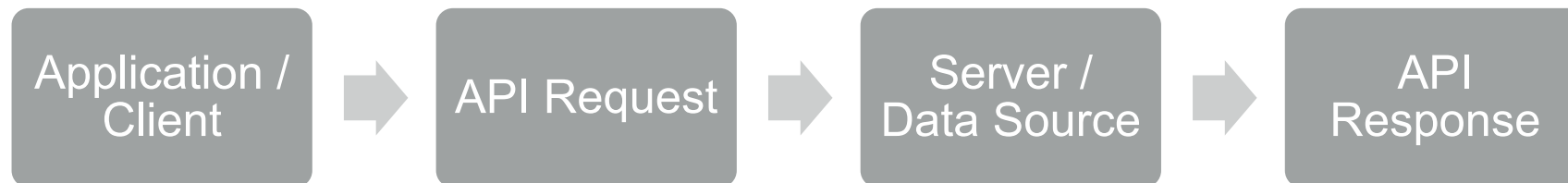


API

Stands for **Application Programming Interface**



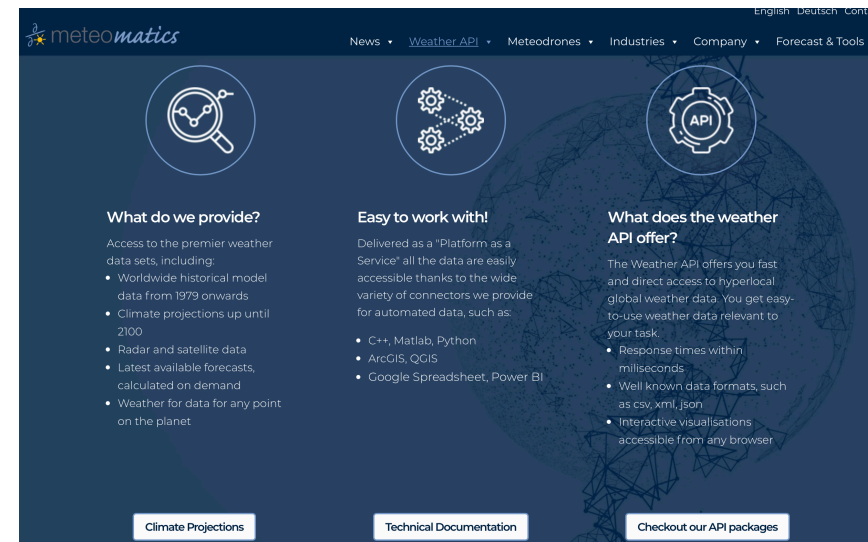
- Written code providing users access to software
- What people use to create interactions with external systems



APIs as Products

Weather Underground packaged its API into a product, selling access to its weather data through the API

https://www.meteomatics.com/en/weather-api/?gclid=Cj0KCQjwuL_8BRCXARIsAGiC51BEYx9zvaM0DPFSwBkDLvu_7u1DzTZkdkBHCUEvQAKlkn-bC7ECcaAubqEALw_wcB



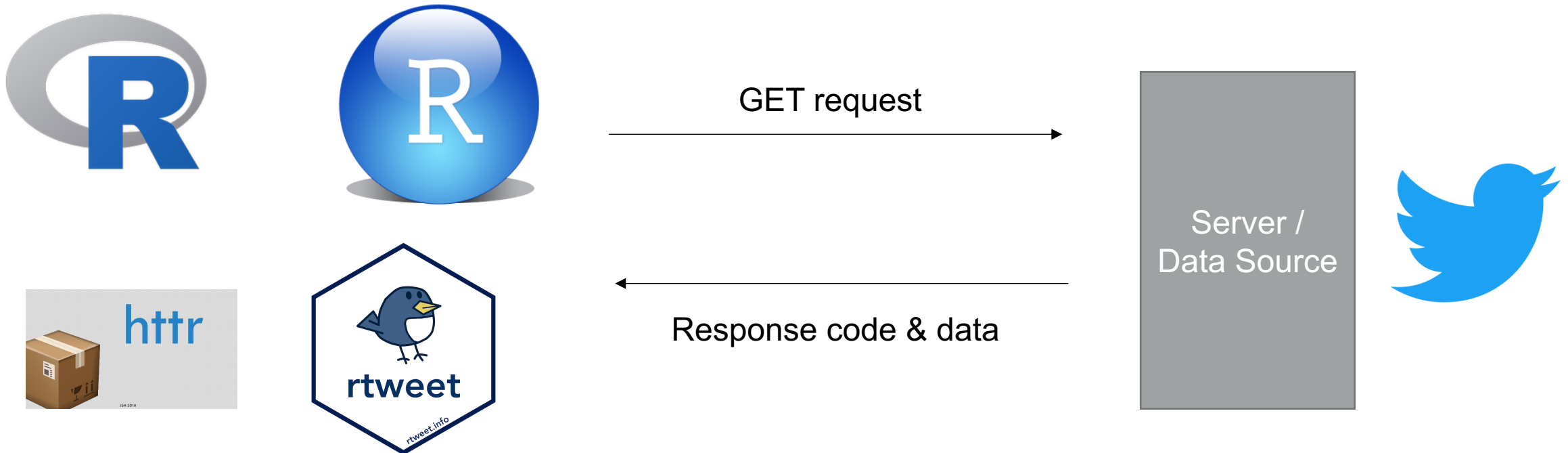
Web Scraping APIs in R

- Register an application, generate consumer keys as well as an access token and an access secret
- Use your application to create an authorization token and to send API requests for data
- Transform and store response data in tables



Web Scraping APIs – R & Twitter

We are the client using R packages as the applications to send out HTTP requests to Twitter's server. The raw API responses we get from Twitter are usually in JSON or XML but one of the packages we will use (rtweet) transforms these files into data frames, making our lives easier



Scraping Twitter Data

```
library(httr)
library(rtweet)

twitter_endpts <- oauth_endpoints("twitter")
twitter_endpts
```

```
twitter_key <- ""
twitter_secret <- ""
access_token <- ""
access_secret <- ""
```

Put information from
your Twitter app here

```
twitter_token <- create_token(app = "twitter",
                             consumer_key = twitter_key,
                             consumer_secret = twitter_secret,
                             access_token = access_token,
                             access_secret = access_secret)
```

Connect with
Twitter here



Scraping Twitter Data

```
# Randomly sample (approximately 1%) from the live stream of all tweets for 10 secs
sample <- stream_tweets(
  q = "",
  timeout = 10)

# Stream Tweets that contain specific keywords for 30 seconds
big4 <- stream_tweets(
  q = "EY, PwC, Deloitte, KPMG",
  timeout = 30)

# Search for Twitter statuses containing a keyword, phrase, or multiple keywords.
# ONLY RETURNS DATA FROM THE PAST 6-9 DAYS.
# To return more than 18,000 statuses in a single call, set "retryonratelimit" to TRUE.
# Default is to return 100 tweets - to return more, set n to a higher value
# search for a keyword
deloitte_en_tweets <- search_tweets("Deloitte", lang = "en", include_rts = FALSE, n = 5000)
save_as_csv(deloitte_en_tweets, "deloitte_en_tweets.csv", prepend_ids = TRUE, na = "", fileEncoding = "UTF-8")
```



Scraping Twitter Data

```
# Search for the timeline of users
# Provide a user ID or screen name and specify the number of tweets (max of 3,200).
deloitte_timeline <- get_timeline("@Deloitte", n = 3200)

# Get the list of account followed by @Deloitte
# i.e., who does @Deloitte follow?
deloitte_friends <- get_friends("@Deloitte")

# Lookup function and Twitter trends
deloitte_following <- lookup_users(deloitte_friends$user_id)

## Store WOEID for Worldwide trends
trends <- trends_available()
trends
atlanta <- trends$woeid[grep("Atlanta", trends$name, ignore.case = TRUE)]
atlanta_trends <- get_trends(atlanta)
```



Ready to scrape some Twitter data?



In-Class Exercise

- Open a Rmd file
- Copy and paste code from slides 8 to 10
 - Change the code to add the information from you Twitter app (if your account has not been approved yet, try using the information from rtweet (i.e., ignore code in slide 8 except for packages))
- Execute code in Markdown
- Open the deloitte_en_tweets.csv file you saved. How much data do you have?

```
1 ---
2 title: "Scraping Twitter Data"
3 author: "Carolina Alves de Lima Salge"
4 date: "2/1/2021"
5 output: word_document
6 ---
7
8 ```{r setup, include=FALSE}
9 knitr::opts_chunk$set(echo = TRUE)
10 ```
11
12 ## Twitter
13
14 Below I will execute code to scrape data from Twitter without putting information from my
15 app. Rather, I will rely on the connection from the `r rtweet` package.
16
17 ```{r}
18 library(httr)
19 library(rtweet)
20
21 # Randomly sample (approximately 1%) from the live stream of all tweets
22 # for 10 secs
23 sample <- stream_tweets(
24   q = "",
25   timeout = 10
26 )
27
28 # Stream Tweets that contain specific keywords for 30 seconds
29 big4 <- stream_tweets(
30   q = "EY, PwC, Deloitte, KPMG",
31   timeout = 30
32 )
33
34 # Search for Twitter statuses containing a key
35 # ONLY RETURNS DATA FROM THE PAST 6-9 DAYS.
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51 ## Store WOEID for Worldwide trends
52 trends <- trends_available()
53 trends
54 atlanta <- trends$woeid[grep("Atlanta", trends$name, ignore.case = TRUE)]
55 atlanta_trends <- get_trends(atlanta)
56
57 ```
```

Data	
atlanta_trends	50 obs. of 9 variables
big4	158 obs. of 90 variables
deloitte_timeline	3198 obs. of 90 variables
deloitte_en_tweets	5000 obs. of 90 variables
deloitte_friends	1463 obs. of 2 variables
deloitte_followi	1462 obs. of 90 variables
sample	480 obs. of 90 variables
trends	467 obs. of 8 variables
Values	
atlanta	2357024L

name	url	parentid
Worldwide	http://where.yahooapis.com/v1/place/1	0
Winnipeg	http://where.yahooapis.com/v1/place/2972	23424775
Ottawa	http://where.yahooapis.com/v1/place/3369	23424775
Quebec	http://where.yahooapis.com/v1/place/3444	23424775
Montreal	http://where.yahooapis.com/v1/place/3534	23424775
Toronto	http://where.yahooapis.com/v1/place/4118	23424775
Edmonton	http://where.yahooapis.com/v1/place/8676	23424775
Calgary	http://where.yahooapis.com/v1/place/8775	23424775
Vancouver	http://where.yahooapis.com/v1/place/9807	23424775
Birmingham	http://where.yahooapis.com/v1/place/12723	23424975

1-10 of 467 rows | 1-3 of 8 columns Previous 1 2 3 4 5 6 ... 47 Next

In-Class Exercise

- Open a Rmd file
- Copy and paste code from slides 8 to 10
 - Change the code to add the information from your Twitter app (if your account has not been approved yet, try using the information from rtweet (i.e., ignore code in slide 8 except for packages))
- Execute code in Markdown
- Open the `deloitte_en_tweets.csv` file you saved. How much data do you have?

Authorize rstats2twitter to access your account?



rstats2twitter
rtweet.info

rstats2twitter is the official app used by rtweet, an open source package/library, to enable collecting and analyzing Twitter data from the REST and stream APIs all while working in the R environment.

Authorize app

Cancel

This application will be able to:

- See Tweets from your timeline (including protected Tweets) as well as your Lists and collections.
- See your Twitter profile information and account settings.
- See accounts you follow, mute, and block.
- Follow and unfollow accounts for you.
- Update your profile and account settings.
- Post and delete Tweets for you, and engage with Tweets posted by others (Like, un-Like, or reply to a Tweet, Retweet, etc.) for you.
- Create, manage, and delete Lists and collections for you.
- Mute, block, and report accounts for you.
- Send Direct Messages for you and read, manage, and delete your Direct Messages.

Learn more about third-party app permissions in the [Help Center](#).

Authentication complete. Please close this page and return to R.



Thank You!

