

Analyzing twitter data

ANALYZING SOCIAL MEDIA DATA IN R



Sowmya Vivek
Data Science Coach

Course Overview

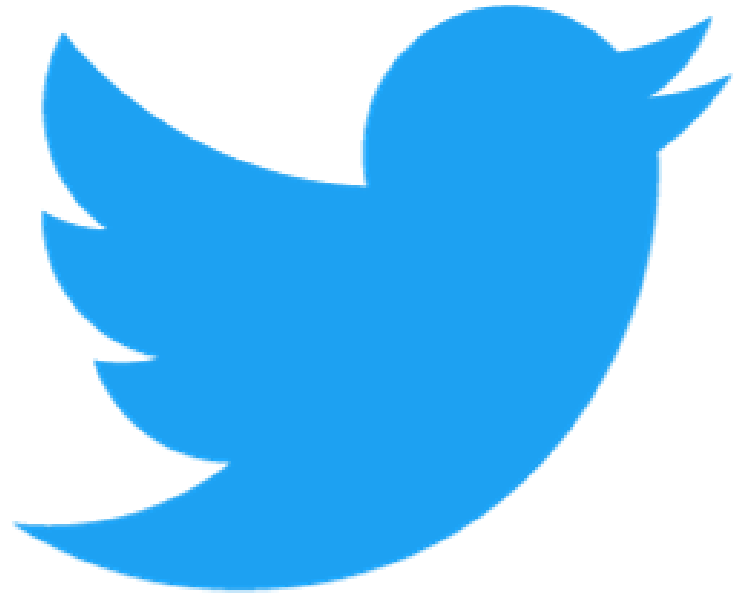
- Extract and visualize twitter data
- Analyze tweet text
- Perform network analysis
- View tweets on the map
- Explore tweets on celebrities, brands, hot topics, and sports

Introduction to social media analysis

- Collect data from social media websites
- Analyze data to derive insights
- Make improved business decisions

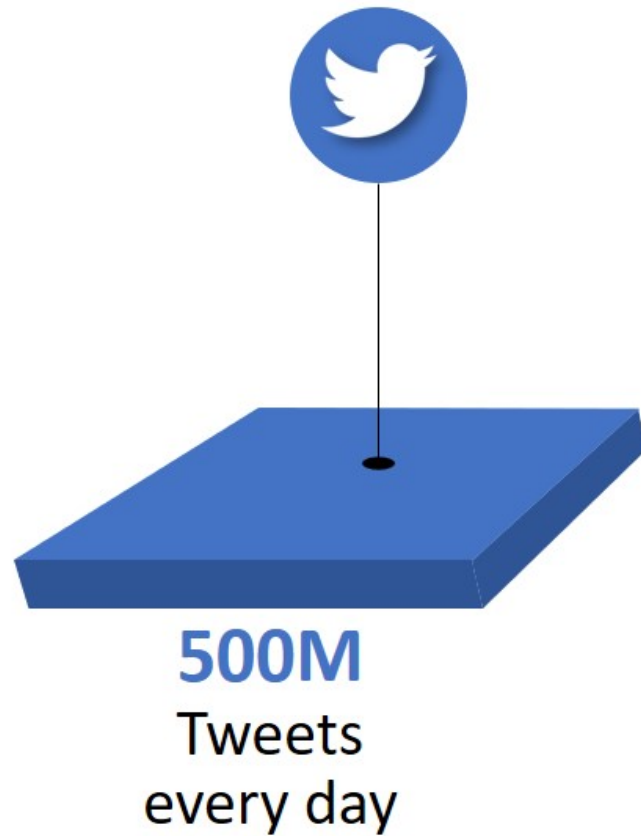


About Twitter

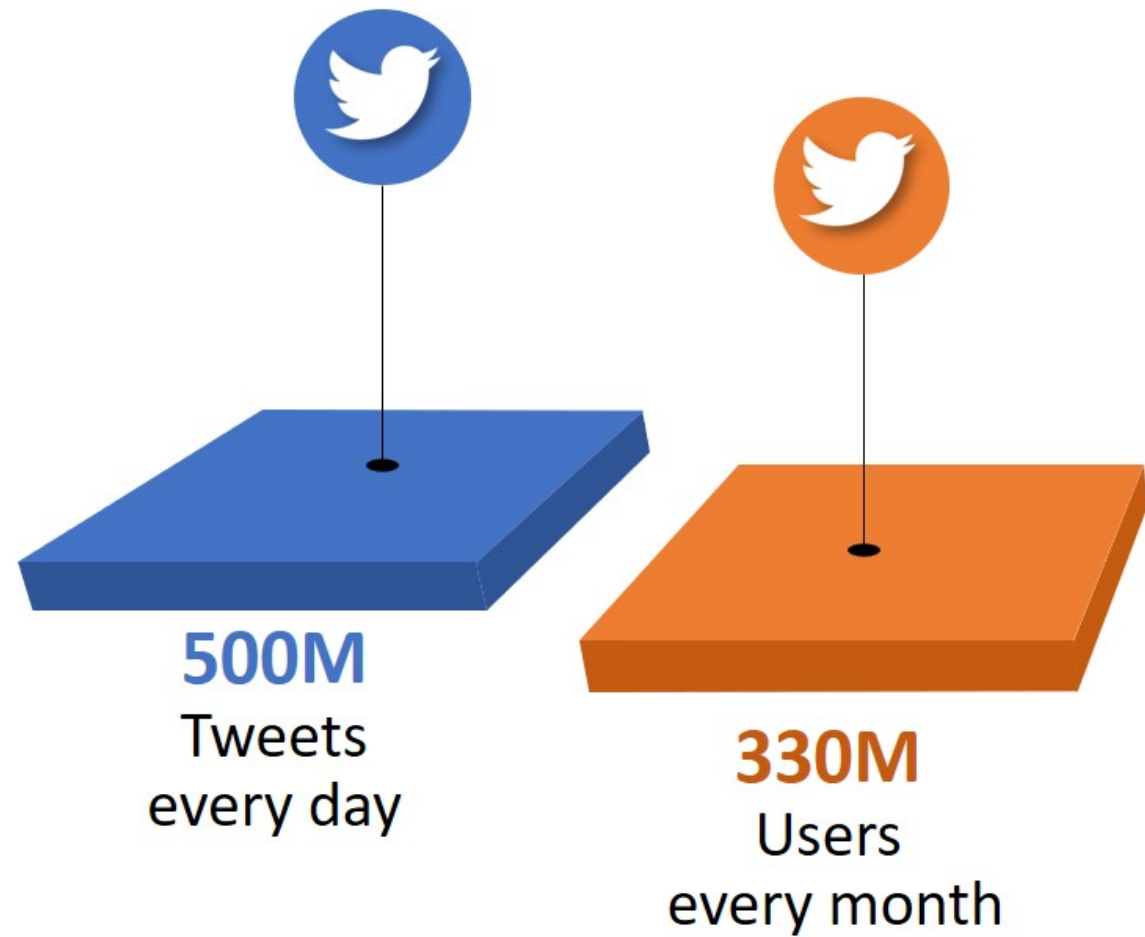


- Social media platform
- Short messages called tweets
- Micro-blogging site
- Information from tweets & tweet metadata

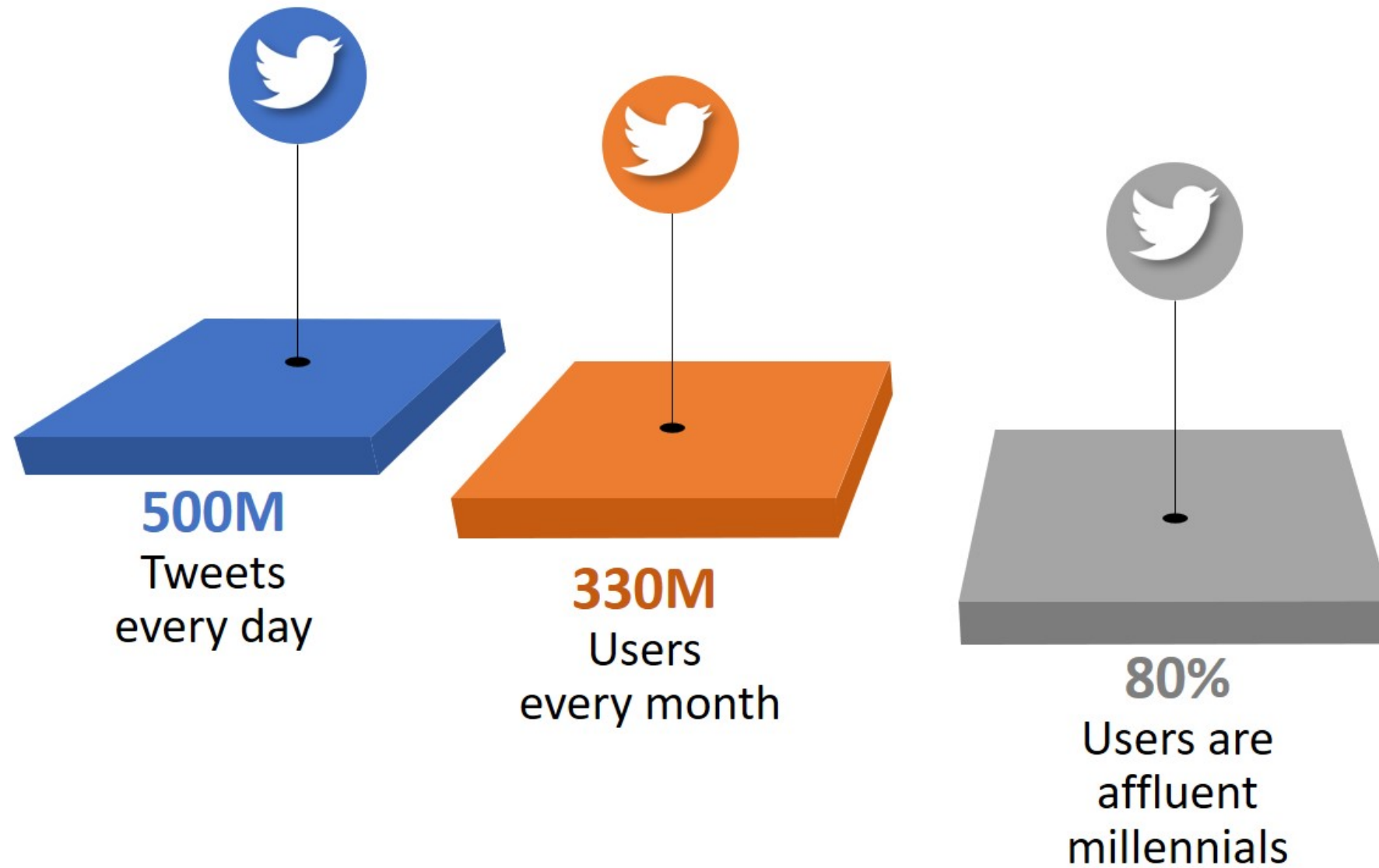
Power of twitter data



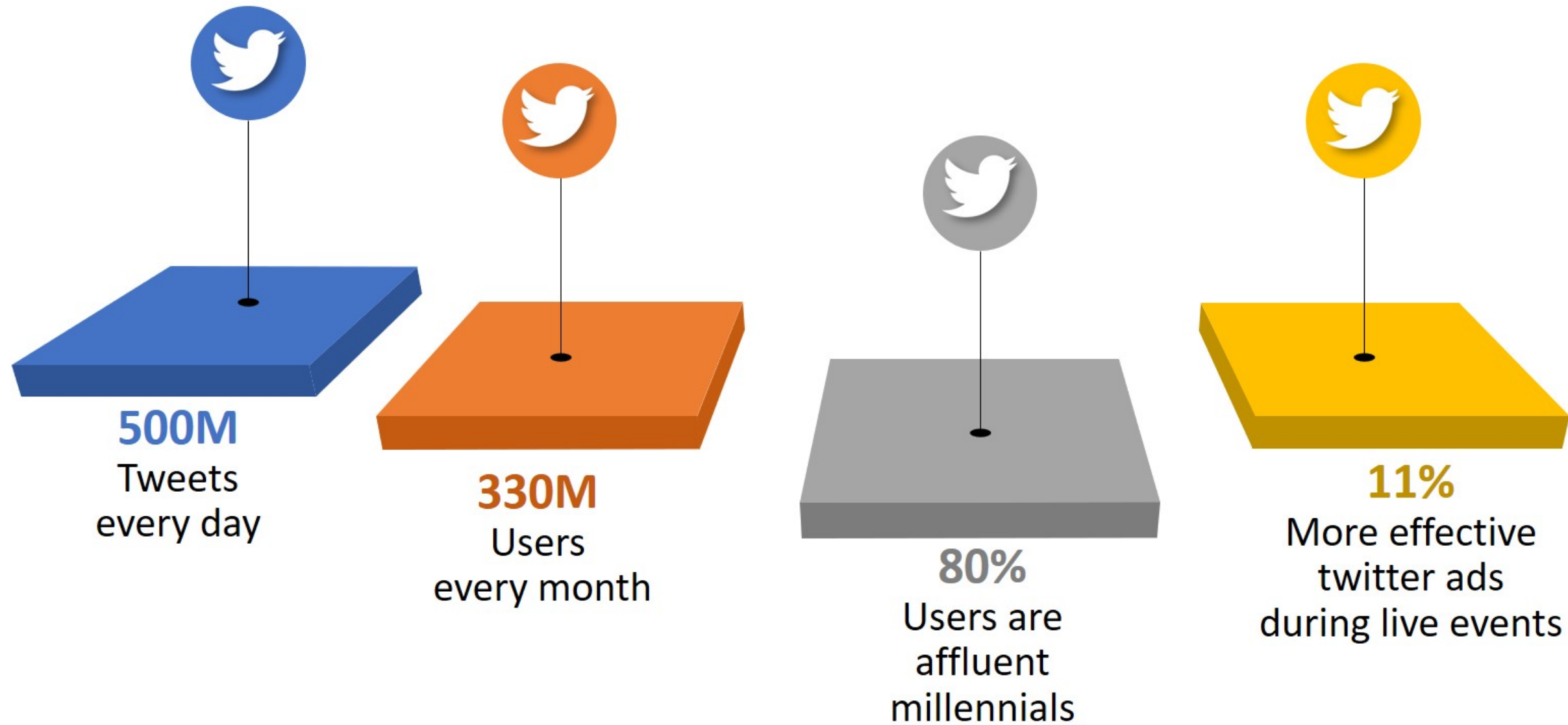
Power of twitter data



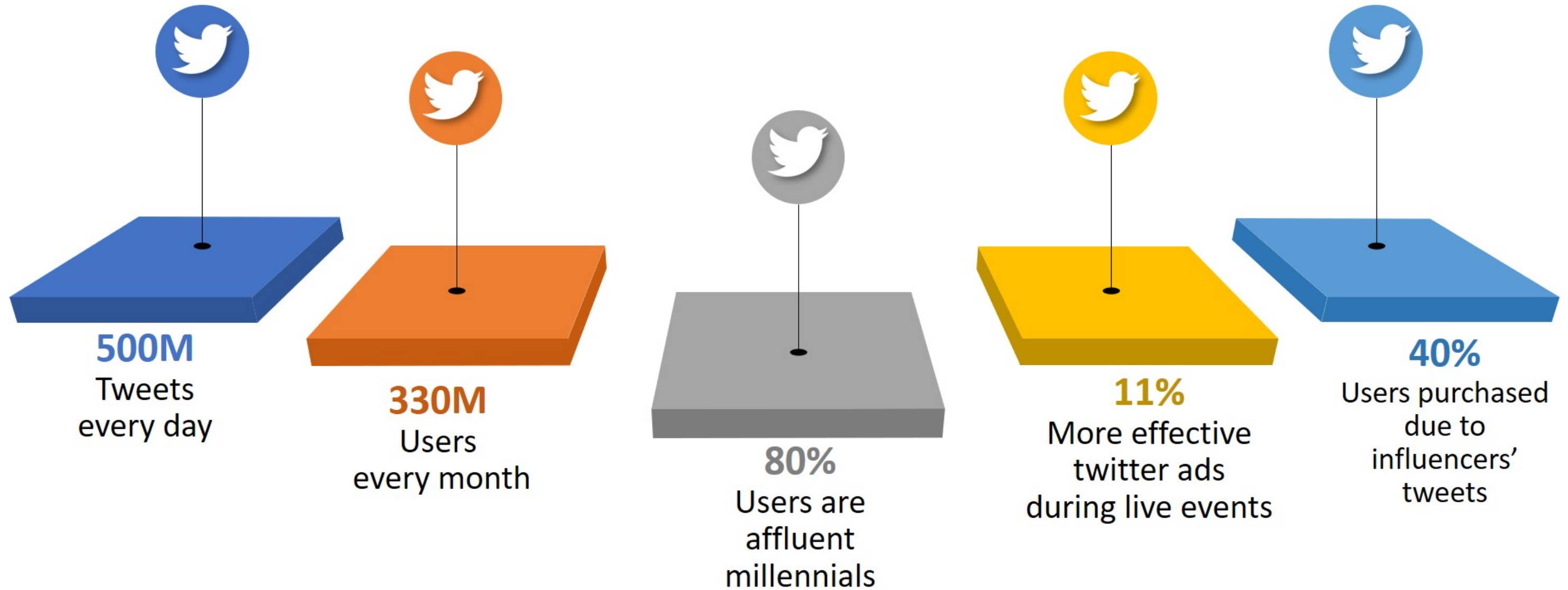
Power of twitter data



Power of twitter data



Power of twitter data



Volume of tweets

- Many functions available in R to extract tweets for analysis
- `stream_tweets()` samples 1% of all publicly available tweets
- Tweets extracted for a 30 second time interval by default

Volume of tweets

```
live_tweets <- stream_tweets("")  
dim(live_tweets)
```

```
[1] 1047  90
```

Volume of tweets

```
live_tweets60 <- stream_tweets("", timeout = 60)  
dim(live_tweets60)
```

```
[1] 3464  90
```

Applications of twitter data

Current trending
topics across the
world

**TRENDING
TOPICS**

Applications of twitter data

Current trending
topics across the
world

**TRENDING
TOPICS**

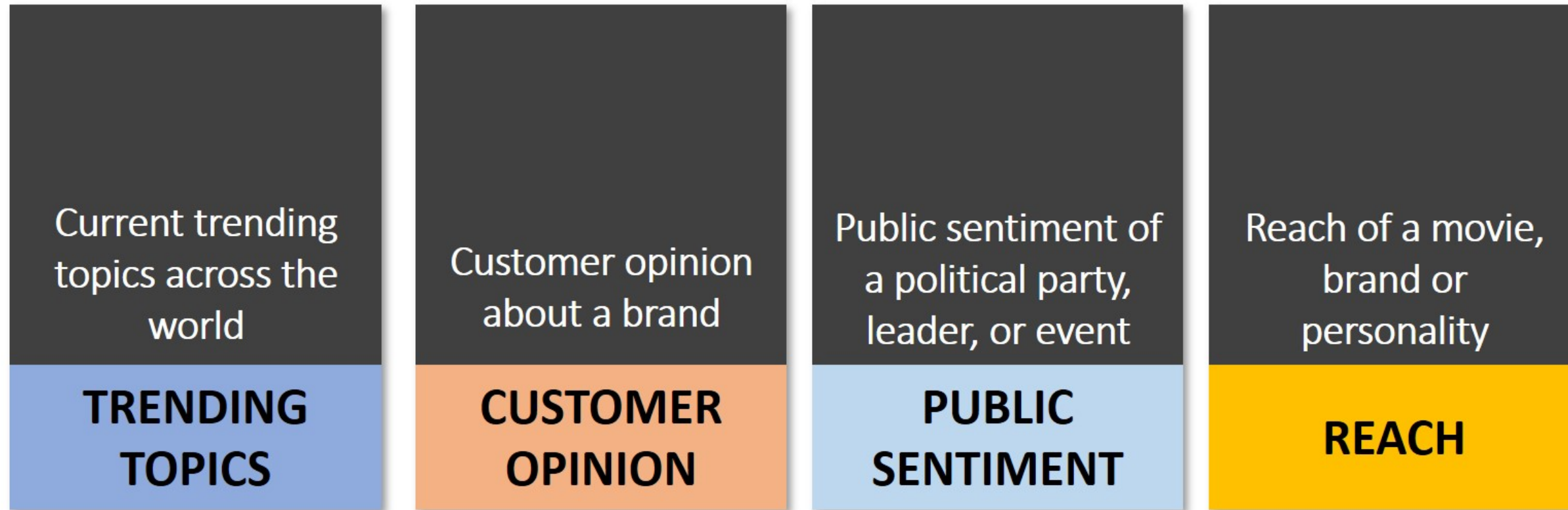
Customer opinion
about a brand

**CUSTOMER
OPINION**

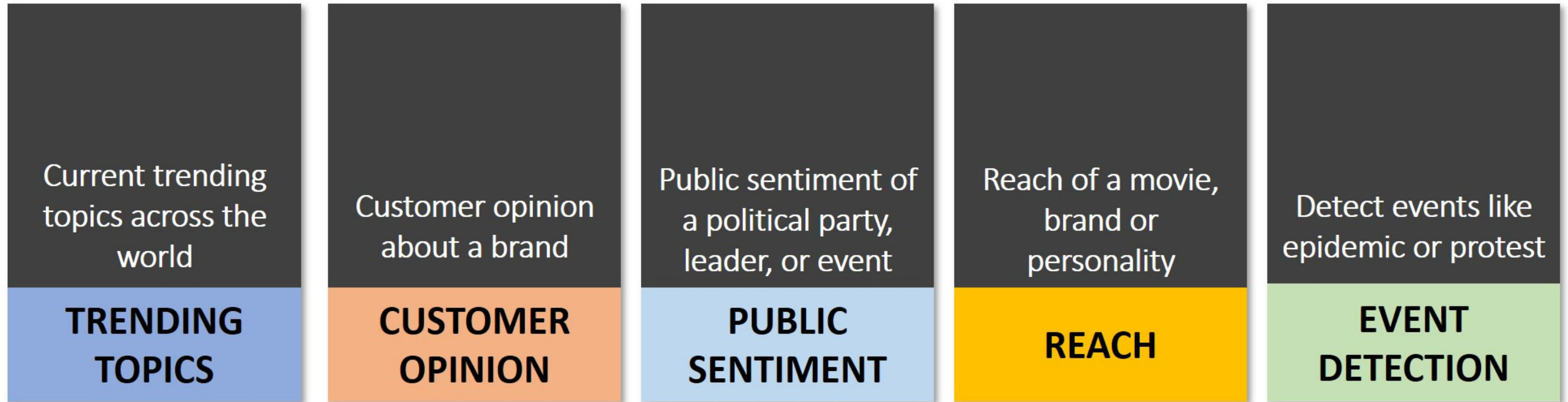
Applications of twitter data



Applications of twitter data



Applications of twitter data



Advantages of twitter data

- Twitter API is open and accessible
- Easier to find conversations because of the hashtag norms
- Since the length of tweets is limited, running algorithms is easy and controlled

Limitations of twitter data

- Historical search is limited for a free account
- A limited number of tweets extracted for a free account
- 1% sample tweets extracted may not be accurate
- Very small % of tweets have geographic tagging

Let's practice!

ANALYZING SOCIAL MEDIA DATA IN R

Extracting twitter data

ANALYZING SOCIAL MEDIA DATA IN R



Sowmya Vivek
Data Science Coach

Lesson Overview

- API fundamentals
- Twitter API types
- Setup the R environment
- Extract data from twitter

API explained

- Application Programming Interface
- Software intermediary that allows two applications to talk to each other
- Twitter APIs interact with twitter and help access tweets

API-based subscriptions

Standard APIs

- Free account
- Basic queries for searching and streaming tweets
- Access to last 7 days tweet data

API-based subscriptions

Standard APIs

- Free account
- Basic queries for searching and streaming tweets
- Access to last 7 days tweet data

Premium and Enterprise APIs

- Paid subscription models
- Access to last 30 days or full archive of tweets

Prerequisites to set up R

- Prerequisites to set up R in your computer
 - A twitter account
 - Pop-up blocker disabled in the browser
 - Interactive R session
 - `rtweet` and `httpuv` packages installed in R
- All prerequisites have been setup within the DataCamp interface

The rtweet and httpuv packages

rtweet

- R package used for extracting data from Twitter API
- Converts twitter data to user friendly data structures

httpuv

- Helps authenticate twitter API access via web browser
- Building block for other R packages

Setting up the R environment

- Steps to set up the R environment in your computer
 - `rtweet` and `httpuv` libraries activated
 - `search_tweets()` function with a search query to connect with twitter
 - Authorize access via browser pop-up
 - "Authentication complete" confirms authorization of twitter access
- R environment has already been setup within the DataCamp interface

Extract twitter data: search_tweets()

- `search_tweets()` returns twitter data matching a search query
- Tweets from the past 7 days only
- Maximum of 18,000 tweets returned per request

```
# Load the rtweet library  
library(rtweet)
```

```
# Extract tweets on "#gameofthrones" using search_tweets()  
tweets_got <- search_tweets("#gameofthrones", n = 1000, include_rts = TRUE, lang = "en")
```

Extract twitter data: search_tweets()

```
head(tweets_got, 4)
```

user_id <chr>	status_id <chr>	created_at <S3: POSIXct>	screen_name <chr>	text <chr>
727816588171350017	1176103860554915841	2019-09-23 11:59:45	LeonardoUzcat1	Today.\n\n#GameofThrones has wo
363838927	1176103859464396806	2019-09-23 11:59:45	mariaaa_carmen	We break the wheel together.\n\
881880538461618176	1176103856163434497	2019-09-23 11:59:44	_valkyriez	The #Emmys had their chance wit
521127287	1176103856075431936	2019-09-23 11:59:44	Nudeus	Congrats to #GameofThrones (60%

Extract twitter data: get_timeline()

- `get_timeline()` extracts tweets posted by a specific twitter user
- Returns upto 3200 tweets

```
# Extract tweets of Katy Perry using get_timeline()  
gt_katy <- get_timeline("@katyperry", n = 3200)
```

Extract twitter data: get_timeline()

```
# View the output  
head(gt_katy)
```

user_id	status_id	created_at	screen_name	text
<chr>	<chr>	<S3: POSIXct>	<chr>	<chr>
21447363	1175132444103565312	2019-09-20 19:39:42	katyperry	My baby angel @cynthialovely
21447363	1175033932355649536	2019-09-20 13:08:15	katyperry	CHICAGO! I'm going to make it
21447363	1174461907656273920	2019-09-18 23:15:13	katyperry	I still dress like a child to
21447363	1174428616735756288	2019-09-18 21:02:56	katyperry	watch me perform ????Small T
21447363	1174381476227338240	2019-09-18 17:55:37	katyperry	???? #SmallTalk ???? with my
21447363	1174061536580497409	2019-09-17 20:44:17	katyperry	Make a ???? connection with

Let's practice!

ANALYZING SOCIAL MEDIA DATA IN R

Components of twitter data

ANALYZING SOCIAL MEDIA DATA IN R



Sowmya Vivek
Data Science Coach

Lesson Overview

- Introduction to twitter JSON
- Extract components of metadata from the JSON
- Use components to derive insights

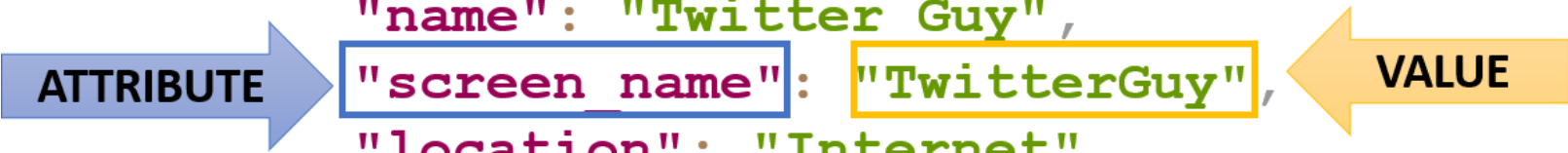
Twitter JSON

- A tweet can have over 150 metadata components
- Tweets and their components returned as JavaScript Object Notation

JSON attributes and values

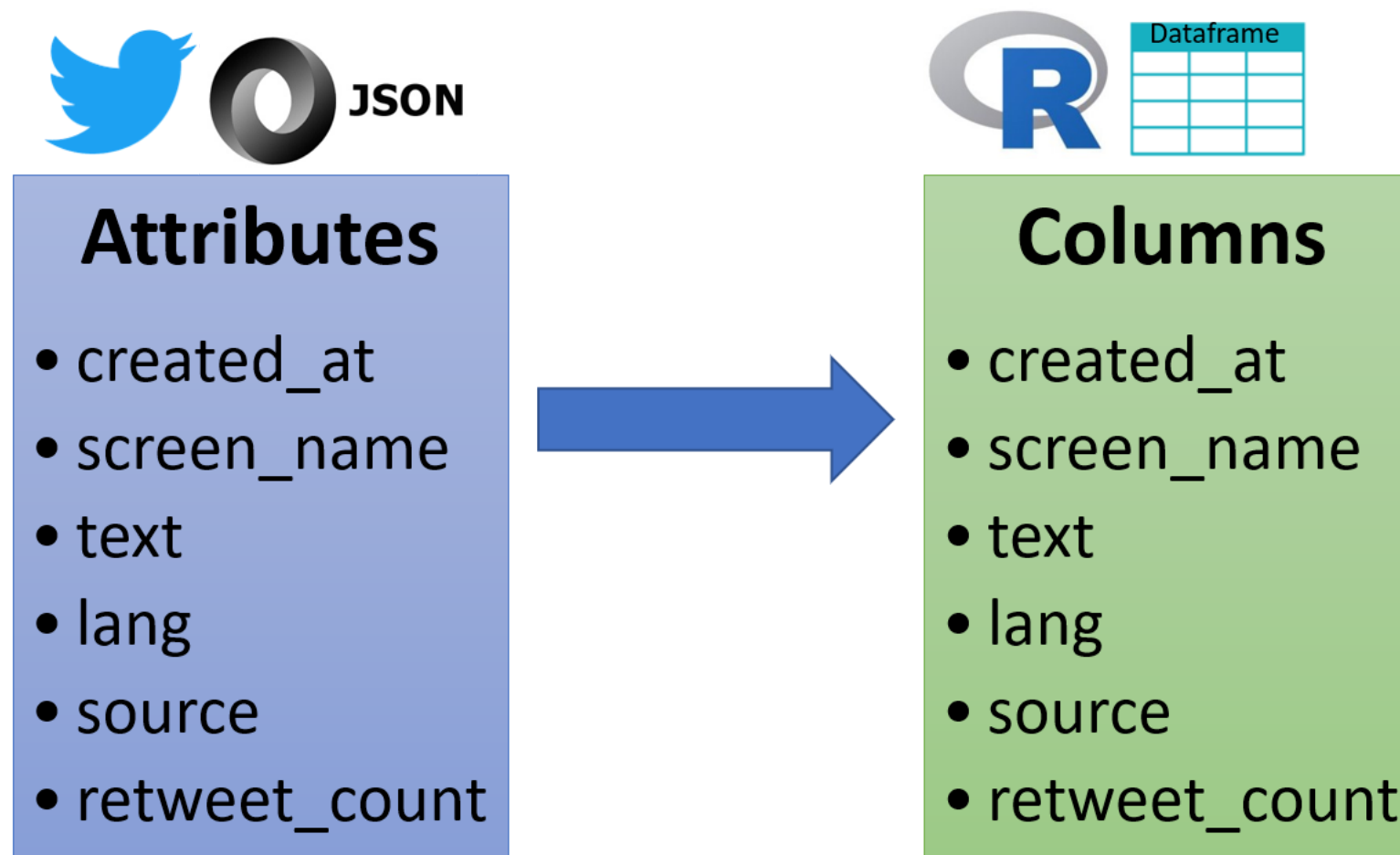
- Attributes and values to describe tweets and components
- Example: `screen_name` stores the twitter handle of a user

```
{  
  "created_at": "Tue Oct 01 02:42:56 +0000 2019",  
  "id_str": "1178862634122317824",  
  "text": "Stock #Options Are Now Bracing for #Brexit or #Trump-Like Shock",  
  "user": {  
    "id": 2244994945,  
    "name": "Twitter Guy",  
    "screen_name": "TwitterGuy",  
    "location": "Internet",  
    "url": "https://\dev.twitter.com/",  
    "description": "Your official source for Twitter Platform news, updates & events"  
  },  
}
```



Converting JSON to a dataframe

- Twitter JSON converted to dataframe by `rtweet` library
- Attributes and values converted to column names and values



Viewing components of tweets

```
# Extract tweets on "#brexit" using search_tweets()  
tweets_df <- search_tweets("#brexit")
```

```
# View the column names  
names(tweets_df)
```

Viewing components of tweets

[1]	"user_id"	"status_id"	"created_at"
[4]	"screen_name"	"text"	"source"
[7]	"display_text_width"	"reply_to_status_id"	"reply_to_user_id"
[10]	"reply_to_screen_name"	"is_quote"	"is_retweet"
[13]	"favorite_count"	"retweet_count"	"quote_count"
[16]	"reply_count"	"hashtags"	"symbols"
[19]	"urls_url"	"urls_t.co"	"urls_expanded_url"
[22]	"media_url"	"media_t.co"	"media_expanded_url"
[25]	"media_type"	"ext_media_url"	"ext_media_t.co"
[28]	"ext_media_expanded_url"	"ext_media_type"	"mentions_user_id"
[31]	"mentions_screen_name"	"lang"	"quoted_status_id"
[34]	"quoted_text"	"quoted_created_at"	"quoted_source"
[37]	"quoted_favorite_count"	"quoted_retweet_count"	"quoted_user_id"
[40]	"quoted_screen_name"	"quoted_name"	"quoted_followers_count"
[43]	"quoted_friends_count"	"quoted_statuses_count"	"quoted_location"
[46]	"quoted_description"	"quoted_verified"	"retweet_status_id"
[49]	"retweet_text"	"retweet_created_at"	"retweet_source"
[52]	"retweet_favorite_count"	"retweet_retweet_count"	"retweet_user_id"
[55]	"retweet_screen_name"	"retweet_name"	"retweet_followers_count"
[58]	"retweet_friends_count"	"retweet_statuses_count"	"retweet_location"
[61]	"retweet_description"	"retweet_verified"	"place_url"
[64]	"place_name"	"place_full_name"	"place_type"
[67]	"country"	"country_code"	"geo_coords"
[70]	"coords_coords"	"bbox_coords"	"status_url"
[73]	"name"	"location"	"description"
[76]	"url"	"protected"	"followers_count"
[79]	"friends_count"	"listed_count"	"statuses_count"
[82]	"favourites_count"	"account_created_at"	"verified"
[85]	"profile_url"	"profile_expanded_url"	"account_lang"
[88]	"profile_banner_url"	"profile_background_url"	"profile_image_url"

Exploring components

- `screen_name` to understand user interest
- `followers_count` to compare social media influence
- `retweet_count` and `text` to identify popular tweets

User interest and tweet counts

- `screen_name` refers to the twitter handle
- Number of tweets posted indicate interest in a topic
- Promote products to interested users

User interest and tweet counts

```
# Extract tweets on "#Arsenal" using search_tweets()  
twts_arsnl <- search_tweets("#Arsenal", n = 18000)
```

```
# Create a table of users and tweet counts for the topic  
sc_name <- table(twts_arsnl$screen_name)  
head(sc_name)
```

```
-----today-----    __JJ23    ___SAbI__    __ambell    __Amzo__    __bobbysingh  
              1              2              3              1              1              1
```

User interest and tweet counts

```
# Sort the table in descending order of tweet counts
sc_name_sort <- sort(sc_name, decreasing = TRUE)
```

```
# View top 6 users and tweet frequencies
head(sc_name_sort)
```

_whatthesport	footy90com	Official_ATG1	TheShortFuse	Rube11M	ArsenalZone_Ind
176	90	88	53	48	43

Followers count

- Count of followers subscribed to a twitter account
- Indicates popularity of the account
- A measure of influence in social media
- Position ads on popular accounts for increased visibility

Compare follower count

```
# Extract user data using lookup_users()  
tvseries <- lookup_users(c("GameOfThrones", "fleabag", "BreakingBad"))
```

```
# Create a dataframe with the columns screen_name and followers_count  
user_df <- tvseries[,c("screen_name", "followers_count")]
```

Compare follower count

```
# View the followers count for comparison
user_df
```

screen_name	followers_count
<chr>	<int>
GameOfThrones	8597188
fleabag	58727
BreakingBad	1240349

Retweet counts and popular tweets

- A retweet is a tweet re-shared by another user
- `retweet_count` stores number of retweets
- Number of retweets helps identify trends
- Popular retweets can be used to promote a brand

Retweet counts and popular tweets

```
# Create a data frame of tweet text and retweet counts  
rtwt <- tweets_arsenal[,c("retweet_count", "text")]
```

```
# Sort data frame based on descending order of retweet counts  
rtwt_sort <- arrange(rtwt, desc(retweet_count))
```

Retweet counts and popular tweets

```
# Exclude rows with duplicate tweet text  
library(data.table)  
rtwt_unique <- unique(rtwt_sort, by = "text")
```

Retweet counts and popular tweets

```
# Print top 6 unique posts retweeted most number of times
head(rtwt_unique)
```

retweet_count	text
<int>	<chr>
5606	Once a Gunner, Always a Gunner. We are proud of you @alexanderiwob
3764	Emirates on Fire ?????????????????? Never give up Gunners?????????????????
2798	That mood tonight ?????? 3?? POINTS ?????? #Arsenal #Gunners #COYG h
2741	#Arsenal fan: "I reckon we'll win the League this season." @Robbie
1687	Auba ?????????????????? This is what I call happiness #aubameyang #arsenal
1166	When sky sports introduced the new Monday night football! The Sha

Let's practice!

ANALYZING SOCIAL MEDIA DATA IN R