Social Media Analytics Workshop Series with R

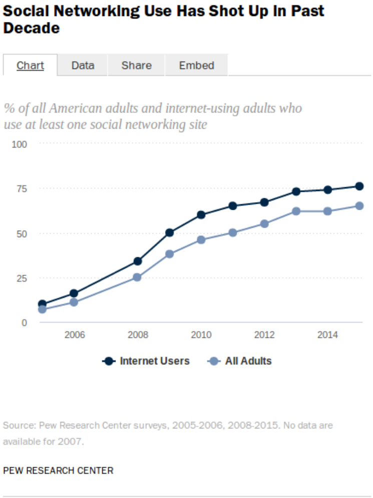
Introduction & Acquiring Social Media Data (Part 1)

Ryan Wesslen
UNC Charlotte / Project Mosaic

July 18, 2017

Why study social media?

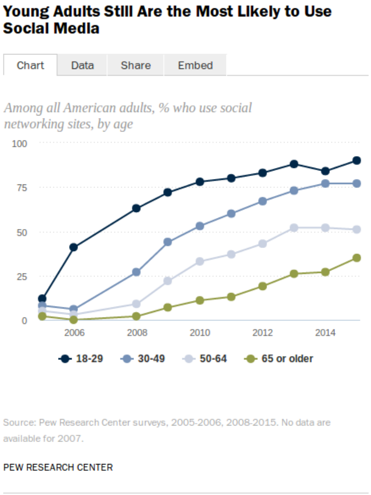
Social Media Use



Social Networking Use Has Shot Up In Past Decade

Source: Pew Research Center surveys, 2005-2006, 2008-2015. No data are available for 2007.

PEW RESEARCH CENTER



Young Adults Still Are the Most Likely to Use Social Media

Source: Pew Research Center surveys, 2005-2006, 2008-2015. No data are available for 2007.

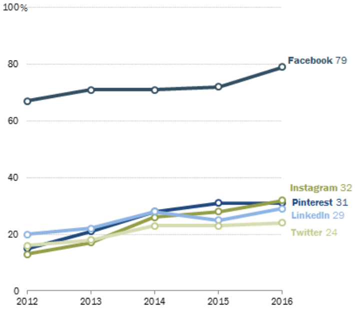
PEW RESEARCH CENTER

`http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/`

Social Media Platform Usage

Facebook remains the most popular social media platform

% of online adults who use ...

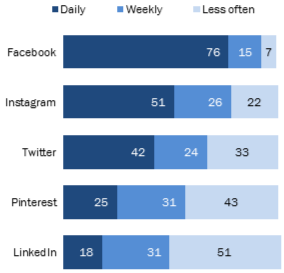


Note: 86% of Americans are currently internet users
Source: Survey conducted March 7-April 4, 2016.
"Social Media Update 2016"

PEW RESEARCH CENTER

Three-quarters of Facebook users and half of Instagram users use each site daily

Among the users of each social networking site, % who use these sites ...

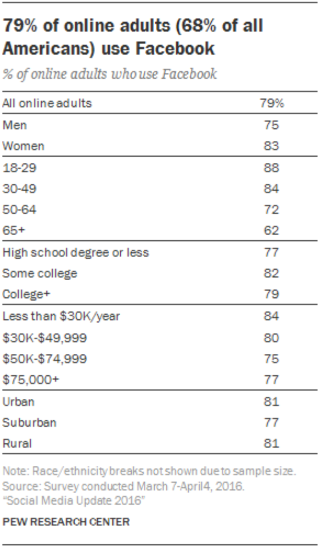


Note: Do not know/refused responses not shown.
Source: Survey conducted March 7-April 4, 2016.
"Social Media Update 2016"

PEW RESEARCH CENTER

http://www.pewinternet.org/2016/11/11/social-media-update-2016/

Representativeness



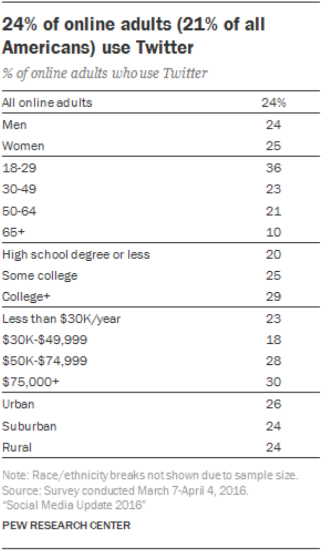
79% of online adults (68% of all Americans) use Facebook

% of online adults who use Facebook

All online adults	79%
Men	75
Women	83
18-29	88
30-49	84
50-64	72
65+	62
High school degree or less	77
Some college	82
College+	79
Less than $30K/year	84
$30K-$49,999	80
$50K-$74,999	75
$75,000+	77
Urban	81
Suburban	77
Rural	81

Note: Race/ethnicity breaks not shown due to sample size.
Source: Survey conducted March 7-April4, 2016.
"Social Media Update 2016"

PEW RESEARCH CENTER



24% of online adults (21% of all Americans) use Twitter

% of online adults who use Twitter

All online adults	24%
Men	24
Women	25
18-29	36
30-49	23
50-64	21
65+	10
High school degree or less	20
Some college	25
College+	29
Less than $30K/year	23
$30K-$49,999	18
$50K-$74,999	28
$75,000+	30
Urban	26
Suburban	24
Rural	24

Note: Race/ethnicity breaks not shown due to sample size.
Source: Survey conducted March 7-April 4, 2016.
"Social Media Update 2016"

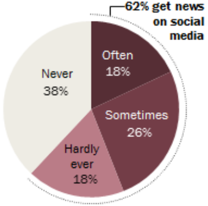
PEW RESEARCH CENTER

http://www.pewinternet.org/2016/11/11/social-media-update-2016/

Social Media as a News Source

About 6-in-10 Americans get news from social media

% of U.S. adults who get news on a social networking site ...

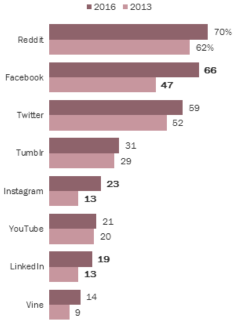


Source: Survey conducted Jan. 12-Feb. 8, 2016.
"News Use Across Social Media Platforms 2016"

PEW RESEARCH CENTER

Growth in use of social media for news

% of users of each social networking site who get news there

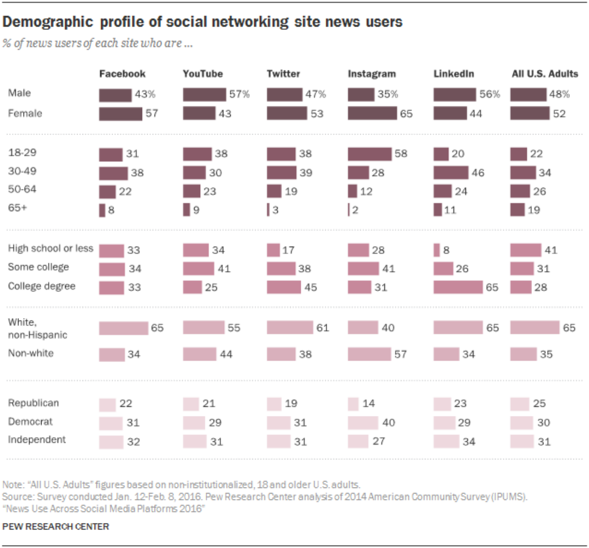


Note: Statistically significant differences in **bold**.
Source: Survey conducted Jan. 12-Feb. 8, 2016.
"News Use Across Social Media Platforms 2016"

PEW RESEARCH CENTER

`http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/`

News Users



Demographic profile of social networking site news users

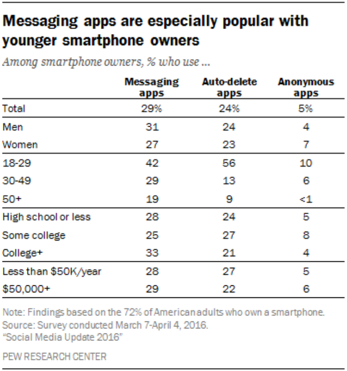
% of news users of each site who are ...

Note: "All U.S. Adults" figures based on non-institutionalized, 18 and older U.S. adults.
Source: Survey conducted Jan. 12-Feb. 8, 2016. Pew Research Center analysis of 2014 American Community Survey (IPUMS).
"News Use Across Social Media Platforms 2016"

PEW RESEARCH CENTER

http://www.journalism.org/2016/05/26/
news-use-across-social-media-platforms-2016/

Growth in Messaging Apps



Messaging apps are especially popular with younger smartphone owners

Among smartphone owners, % who use ...

	Messaging apps	Auto-delete apps	Anonymous apps
Total	29%	24%	5%
Men	31	24	4
Women	27	23	7
18-29	42	56	10
30-49	29	13	6
50+	19	9	<1
High school or less	28	24	5
Some college	25	27	8
College+	33	21	4
Less than $50K/year	28	27	5
$50,000+	29	22	6

Note: Findings based on the 72% of American adults who own a smartphone.
Source: Survey conducted March 7-April 4, 2016.
"Social Media Update 2016"

PEW RESEARCH CENTER

Examples:

- Messaging Apps: WhatsApp, Kik
- Auto-Delete Apps: SnapChat, Wickr
- Anonymous Apps: YikYak, Whisper

`http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/`

Open Questions, Challenges and Ethics for Social Media Data

Big Questions for Social Media Big Data

Four Methodological Considerations

1. Twitter as Model Organism
2. Hashtag Analyses / Selecting on Dependent Variable
3. Missing Denominator
4. Missing the Ecology for the Platform

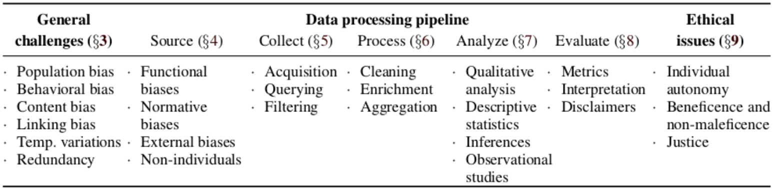
General Practical Steps

1. Target non-social media dependent variables (e.g., election results, employment)
2. Qualitative Pull-outs (validate, validate, validate)
3. Baseline panels
4. Convergent answers and complimentary methods
5. Multi-disciplinary Teams

Source: Tufekci, 2014 `https://arxiv.org/pdf/1403.7400.pdf`

Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries

Table I: General challenges along an idealized data processing pipeline and ethical considerations.



General	**Data processing pipeline**					**Ethical**
challenges (§3)	Source (§4)	Collect (§5)	Process (§6)	Analyze (§7)	Evaluate (§8)	**issues (§9)**
· Population bias · Behavioral bias · Content bias · Linking bias · Temp. variations · Redundancy	· Functional biases · Normative biases · External biases · Non-individuals	· Acquisition · Querying · Filtering	· Cleaning · Enrichment · Aggregation	· Qualitative analysis · Descriptive statistics · Inferences · Observational studies	· Metrics · Interpretation · Disclaimers	· Individual autonomy · Beneficence and non-maleficence · Justice

"For your own sanity, you have to remember that not all problems can be solved. Not all problems can be solved, but all problems can be illuminated." -Ursula Franklin (via M. Meredith)

Olteanu et al., 2017: `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2886526`

Population Biases (Olteanu et al., 2017)

Population Biases: Differences in demographics or other characteristics between a population of users represented in a dataset or platform and a target population.

1. Different user demographics tend to be drawn to different social platforms.
2. Users of different demographics or traits use social platforms mechanisms differently.
3. Proxies for personal traits or demographic criteria vary in reliability.

Example: Geolocation in Twitter

There are three types of geolocation Twitter data:

1. Points: latitude and longitude
2. Places (polygons): bounding boxes associated with place name
3. Profile location: open-ended text

- Only about 1-2.9% of tweets are geocoded as a point or polygon (Graham et al. 2014; Osborne and Dredze 2014).
- About 60% of Twitter users' profiles contain a valid location (Hecht et al. 2011)
- Can also use entity extraction of the body of the tweet (mention locations).

`http://support.gnip.com/articles/geo-intro.html`

Functional Biases (Olteanu et al., 2017)

Functional Biases: Biases that are a result of platform-specific mechanisms or affordances, that is, the possible actions within each system or environment.

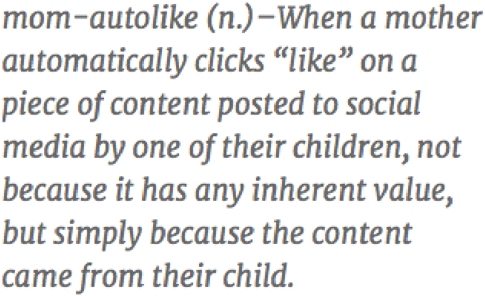
1. Platform-specific design and features shape user behavior.
2. Algorithms used for organizing and ranking content influence user behavior.
3. The way in which contents are presented influences user behavior.

Machine intelligence makes human morals more important: Zeynip Tufekci

“We need to audit our black boxes”

`https://youtu.be/hSSmmlridUM?t=649`

Facebook Mom Algorithm Problem



mom-autolike (n.)–When a mother automatically clicks "like" on a piece of content posted to social media by one of their children, not because it has any inherent value, but simply because the content came from their child.

"The problem is: Facebook, despite the fact that they know she's my mom, doesn't take this fact into account in their algorithm."

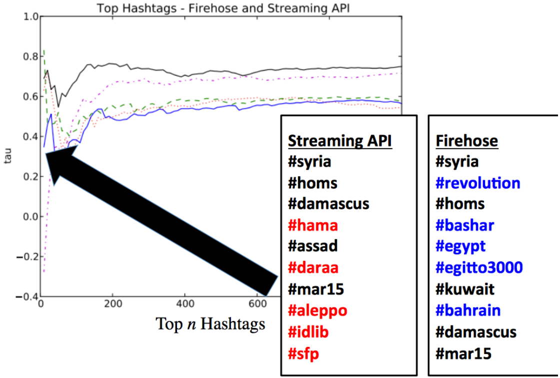
`http://boffosocko.com/2017/07/11/the-facebook-algorithm-mom-problem/`

Data Collection Biases (Olteanu et al., 2017)

Data Collection Biases: Biases introduced due to the selection of data sources, or by the way in which data from these sources are acquired and prepared.

1. Many social platforms discourage data collection by third parties.
2. Programmatic access to data from a platform comes with limitations.
3. The platform may not capture all relevant data.
4. Platforms may not give access to all the data they capture.
5. Sampling strategies are often opaque.

Potential Biases in Twitter API's



Streaming API	Firehose
#syria	#syria
#homs	#revolution
#damascus	#homs
#hama	#bashar
#assad	#egypt
#daraa	#egitto3000
#mar15	#kuwait
#aleppo	#bahrain
#idlib	#damascus
#sfp	#mar15

Fred Morstatter, Jürgen Pfeffer, Huan Liu, Kathleen M Carley. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose, ICWSM 2013.

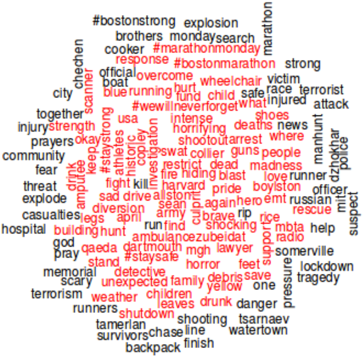
Fred Morstatter, and Huan Liu. Social Data Bias in Machine Learning: Impact, Evaluation, and Correction, AAAI 2017 Tutorial SP6.

Data Querying (Olteanu et al., 2017)

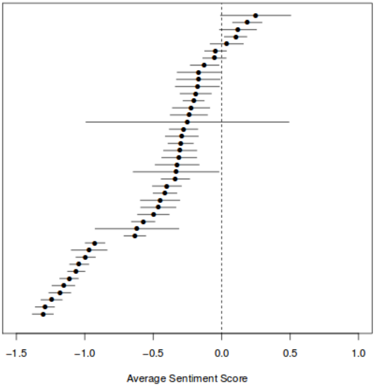
Data Querying for social platforms typically result in a filtering of data expressed through a query through an API.

1. APIs have limited expressiveness regarding information needs.
2. An information need may be operationalized to an API in different ways.
3. The choice of keywords in keyword-based queries shapes the resulting datasets.

Example: Keyword Querying is a non-trivial task


#bostonstrong explosion marathon
brothers monday search
cooker #marathonmonday
chechen response #bostonmarathon strong
official overcome
boat wheelchair victim
city scanner blue running hurt fund child safe race terrorist
#wewillneverforget what injured attack
together usa intense shoes
injury strength horrifying deaths news manhunt
prayers okay shootout arrest where
community #staystrong athletes historic copley investigation swat collier guns people dzhokhar police
fear keep restrict dead madness
threat drink amputee fire hiding blast love runner
fight kill harvard pride boylston officer
explode sad drive allston emt russian mit suspect
casualties legs diversion sean again hero rescue
april army culprit brave rip rice
hospital building hunt run find shocking support mbta help
god ambulance zubeidat radio
pray qaeda dartmouth mgh lawyer somerville
stand #staysafe horror feet
memorial detective lockdown
scary unexpected family debris save pressure tragedy
terrorism yellow one
weather children
runners leaves drunk danger
shutdown shooting tsarnaev
tamerlan
survivors chase line watertown
backpack finish

(b) Boston Bombings



King, Lam, and Roberts (2017) `http://j.mp/2nxUa8N`

Behavioral Biases (Olteanu et al., 2017)

Behavioral Biases: Differences in user behavior across platforms or contexts, or across users represented in different datasets.

Interaction biases affect how users interact with each other.

- Communication is affected by relationships users have and features of the platform.

Content consumption biases affect how users interact with content.

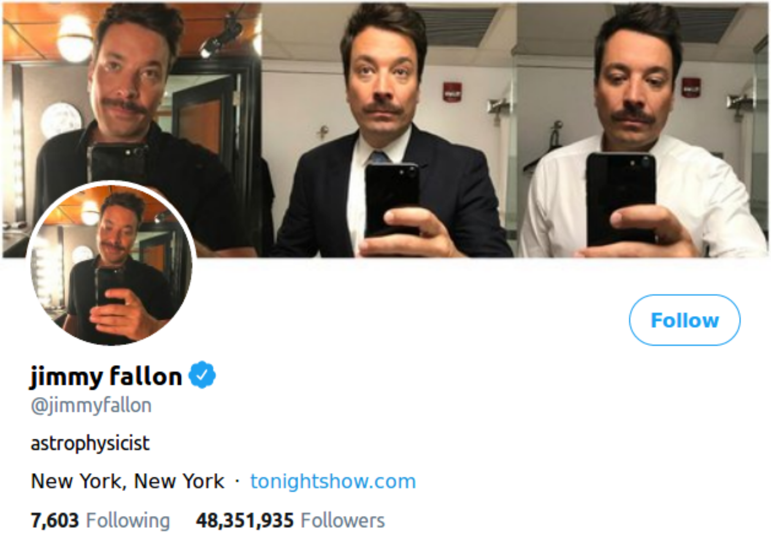
- Users tend to consumer more content from like-minded people; create "filter bubbles" (Nikolov et al. 2015)

Nature of users' tasks influences the traces they leave.

Misreports and self-selection may occur due to behavioral biases.

- Self reported data (e.g., Twitter profile information) may be subject to bias as it effects what/how users chose to provide information

Example: Truthfulness



jimmy fallon ✓

@jimmyfallon

astrophysicist

New York, New York · tonightshow.com

7,603 Following 48,351,935 Followers

Non-Individual Accounts (Olteanu et al., 2017)

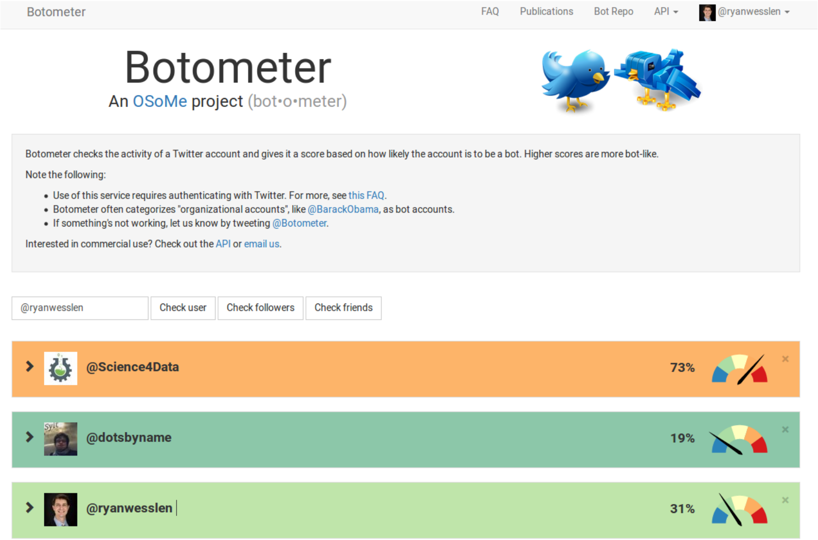
Non-individual agents: Interactions on social platforms that are not produced by individuals, but by accounts representing various types of organizations, or by automated agents.

1. Organizations (e.g., NGOs, government, businesses, media)
2. Automated Agents - bots or spammers

The Rise of Social Bots By Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, Alessandro Flammini Communications of the ACM, Vol. 59 No. 7, Pages 96-104 10.1145/2818717

`https://cacm.acm.org/magazines/2016/7/204021-the-rise-of-social-bots/fulltext`

Example: Bots on bots on bots



Botometer

An OSoMe project (bot•o•meter)

Botometer checks the activity of a Twitter account and gives it a score based on how likely the account is to be a bot. Higher scores are more bot-like.

Note the following:

- Use of this service requires authenticating with Twitter. For more, see this FAQ.
- Botometer often categorizes "organizational accounts", like @BarackObama, as bot accounts.
- If something's not working, let us know by tweeting @Botometer.

Interested in commercial use? Check out the API or email us.

@Science4Data	73%
@dotsbyname	19%
@ryanwesslen	31%

OSoMe project Indiana University `https://botometer.iuni.iu.edu/`

Ethical Considerations (Olteanu et al., 2017)

Just because social data is often publicly accessible does not mean research done on it is always ethical (boyd & Crawford, 2012; Zimmer 2012).

Social media research is different from clinical trials.

- awareness and manipulation, e.g., lab-based image annotation
- awareness without manipulation, e.g., opt-in nutrition and exercise monitoring
- no awareness with manipulation, e.g., A/B testing of features on social media
- no awareness and no manipulation, e.g., observational Twitter studies

Consent

- Obtaining consent from million of users is impractical
- Terms of use for social platform may not constitute informed consent
  - Example: Kramer et al. 2014 `http://www.pnas.org/content/111/24/8788.short`

Ways to Collect Social Media Data

Collecting Social Media Data

Two different methods:

1. Web scraping: extract data from source code of website
2. Web APIs (application programming interface): use structured https requests that return JSON or XML files

Two types of API's:

1. RESTful APIs: queries for static information at current moment
2. Streaming APIs: changes in users' data in real time (future)

Potential Issues:

1. Rate limits: restrictions on number of API calls by user/period of time
2. Ongoing debate on replication of using social media data.

Source: Pablo Barberá
`https://github.com/pablobarbera/social-media-workshop`

Facebook API

Facebook allows access to two types of data through the public API:

1. Data from public Facebook pages (posts, likes, comments)
2. User's personal data (profile, checkins, likes. . .)

Some public user data (gender, location) was available through previous versions of the API (not anymore)

Access to other (anonymized) data used in published studies requires permission from Facebook

R library: `Rfacebook`

Source: Pablo Barberá
`https://github.com/pablobarbera/social-media-workshop`

Twitter API's

Two different methods to collect Twitter data:

1. RESTful API: `twitteR`

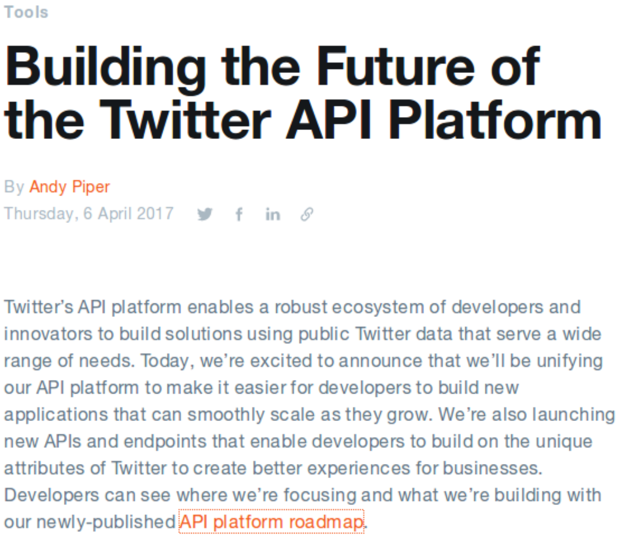
   - Queries for specific information about users and tweets
   - Examples: user profile, list of followers/friends, tweets generated by user's timeline

2. Streaming API: `streamR`

   - Connect to the "stream" of tweets as they're published (future)
   - Three streaming API's:
     1. Filter stream: filtered by keywords, geo, user or language
     2. User stream: filtered by authenticated user (timeline or tweets)
     3. Sample stream: 1% random sample of tweets

Source: Pablo Barberá
`https://github.com/pablobarbera/social-media-workshop`

API’s are constantly changing



Tools

Building the Future of the Twitter API Platform

By Andy Piper

Thursday, 6 April 2017

Twitter’s API platform enables a robust ecosystem of developers and innovators to build solutions using public Twitter data that serve a wide range of needs. Today, we’re excited to announce that we’ll be unifying our API platform to make it easier for developers to build new applications that can smoothly scale as they grow. We’re also launching new APIs and endpoints that enable developers to build on the unique attributes of Twitter to create better experiences for businesses. Developers can see where we’re focusing and what we’re building with our newly-published API platform roadmap.

`https://blog.twitter.com/developer/en_us/topics/tools/2017/building-the-future-of-the-twitter-api-platform.html`